Mellerson, Kendra

Subject: FW: 09890973

-----Original Message-----
From: Gakh, Yelena
Sent: Tuesday, August 05, 2003 2:33 PM
To: STIC-EIC1700
Subject: 09890973

Dear Kendra:

please order one more list:

1. TITLE: "The opportunities and challenges of personalized genome-based molecular therapies for cancer: targets, technologies, and molecular chaperones"
AUTHOR(S): *Workman, Paul*
CORPORATE SOURCE: Cancer Research UK Centre for Cancer Therapeutics, Institute of Cancer Research, Sutton, Surrey, SN2 5NG
SOURCE: **Cancer Chemotherapy and Pharmacology (2003), 52(s01), 45-56**

Thank you,

Yelena

Yelena G. Gakh, Ph.D.

Patent Examiner
USPTO, cp3/7B-08
(703)306-5906

Cancer Chemother Pharmacol (2003) 52 (Suppl 1): S45–S56
DOI 10.1007/s00280-003-0593-0

Paul Workman

The opportunities and challenges of personalized genome-based molecular therapies for cancer: targets, technologies, and molecular chaperones

Published online: 18 June 2003

Abstract There are now unprecedented opportunities for the development of improved drugs for cancer treatment. Following on from the Human Genome Project, the Cancer Genome Project and related activities will define most of the genes in the majority of common human cancers over the next 5 years. This will provide the opportunity to develop a range of drugs targeted to the precise molecular abnormalities that drive various human cancers and opens up the possibility of personalized therapies targeted to the molecular pathology and genomics of individual patients and their malignancies. The new molecular therapies should be more effective and have less-severe side effects than cytotoxic agents. To develop the new generation of molecular cancer therapeutics as rapidly as possible, it is essential to harness the power of a range of new technologies. These include: genomic and proteomic methodologies (particularly gene expression microarrays); robotic high-throughput screening of diverse compound collections, together with *in silico* and fragment-based screening techniques; new structural biology methods for rational drug design (especially high-throughput X-ray crystallography and nuclear magnetic resonance); and advanced chemical technologies, including combinatorial and parallel synthesis. Two major challenges to cancer drug discovery are: (1) the ability to convert potent and selective lead compounds with activity by the desired mechanism on tumor cells in culture into agents with robust, drug-like properties, particularly in terms of pharmacokinetic and metabolic properties; and (2) the development of validated pharmacodynamic endpoints and molecular markers of drug response, ideally using noninvasive imaging technologies. The use of various new technologies will be exemplified. A major conceptual and practical issue facing the development and use of the new molecular cancer therapeutics is whether a single drug that targets one of a series of key molecular abnormalities in a particular cancer (e.g. BRAF) will be sufficient on its own to deliver clinical benefit ("house of cards" and tumor addiction models). The alternative scenario is that it will require either a combination of agents or a class of drug that has downstream effects on a range of oncogenic targets. Inhibitors of the heat-shock protein (HSP) 90 molecular chaperone are of particular interest in the latter regard, because they offer the potential of inhibiting multiple oncogenic pathways and simultaneous blockade of all six "hallmark traits" of cancer through direct interaction with a single molecular drug target. The first-in-class HSP90 inhibitor 17AAG exhibited good activity in animal models and is now showing evidence of molecular and clinical activity in ongoing clinical trials. Novel HSP90 inhibitors are also being sought. The development of HSP90 inhibitors is used to exemplify the application of new technologies in drug discovery against a novel molecular target, and in particular the need for innovative pharmacodynamic endpoints is emphasized as an essential component of hypothesis-testing clinical trials.

Keywords Molecular pathology and genomics of cancer · New molecular targets · Technologies for drug discovery and development · HSP90 molecular chaperone inhibitors · Gefitinib · Imatinib · Trastuzumab

This work was presented at the 18th Bristol-Myers Squibb Nagoya International Cancer Treatment Symposium, "New Strategies for Novel Anticancer Drug Development," 8–9 November 2002, Nagoya, Japan

P. Workman
Cancer Research UK Centre for Cancer Therapeutics,
Institute of Cancer Research, Sutton, Surrey,
SN2 5NG, UK
E-mail: paul.workman@icr.ac.uk
Tel.: +44-20-87224301
Fax: +44-20-86424324

Introduction

In many ways cancer drug discovery is unrecognizable from what it was even as little as 10 years ago. The

progressive elucidation of the molecular control pathways that are hijacked by cancers has provided us with a large number of potential targets for therapeutic intervention. At the same time, the putting together of a powerful tool kit of innovative technologies has allowed us to accelerate the pace and improve the efficiency of drug discovery [44].

Hence the focus of the first part of this commentary is on new targets and technologies. To illustrate how new drug discovery and development is now done, the second part of the article comprises a summary and update on the development of inhibitors of the heat-shock protein (HSP) 90 molecular chaperone. These are of particular interest because they provide a potential approach to block combinatorial oncogenesis within a single drug molecule. In addition, the first-in-class HSP90 inhibitor 17AAG is just completing phase I trials with promising early results.

From cancer genes to individualized therapies

Given that we now understand in increasing detail the molecular abnormalities that drive the process of malignant progression, the major strategy for drug discovery in cancer is to identify the genes and cognate biochemical pathways that are hijacked in cancer cells, to discover molecular reagents and biomarkers to identify pathways with these defects, and to develop drugs that counteract or exploit the deregulated control mechanisms. The vision is that we can exploit our growing knowledge of cancer genes and pathways by developing personalized therapies targeted to the molecular pathology of individual patients and their malignancies (see references 44 and 49, and Fig. 1).

Fig. 1 Strategy for exploiting knowledge of cancer genes and pathways in the development of personalized therapies targeted to molecular pathology of individual patients

A range of drugs that target the molecular pathology of cancer are now undergoing clinical trial (e.g. see reference 49, and Table 1). Proof of concept for the approach is provided by the regulatory approval of imatinib (Gleevec), trastuzumab (Herceptin), and gefitinib (Iressa). Various small-molecule cyclin-dependent kinase inhibitors, e.g. flavopiridol and CYC202 (*R*-roscovitine), are undergoing clinical evaluation. Furthermore, a wide range of innovative agents are in preclinical and clinical development. These include drugs that block the farnesylation of RAS and other protein targets; inhibitors of signal transduction kinases such as RAF-1, MEK, mTOR, and PI3 kinase; and drugs that block chromatin remodeling enzymes such as histone deacetylases [49].

The success with the first initial wave of molecular therapeutics that specifically attack the oncogenic pathways that are hijacked by cancer genome defects has provided encouragement for the view that this represents a major opportunity to develop innovative cancer drugs. Furthermore, the mechanism of action of these agents offers potential not only for improved therapeutic efficacy, but also for less-severe side effects compared with the previous generation of cytotoxic agents. The new agents may in fact be much more like tamoxifen—used chronically for long-term disease control and potentially for chemoprevention.

Additional new targets from cancer genomics

A further tranche of new targets and drugs can be expected to emerge over the next 5–10 years as the genes involved in all stages of the malignant progression of every tumor type are elucidated. Historically, cancer genes have been discovered and cloned by a variety of means, including the dissection of major chromosomal abnormalities, i.e. translocations, amplifications, and deletions; transfection of dominant oncogenes into

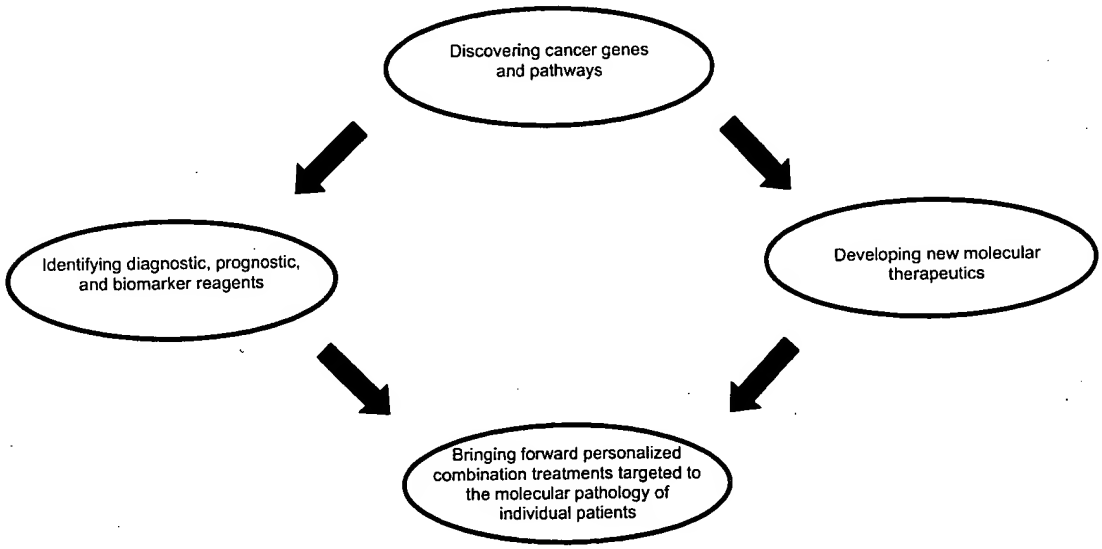


Table 1 Examples of novel drugs acting on cancer genome targets (for further details see reference 49)

Imatinib	A small molecule that shows activity in chronic myeloid leukemia and gastrointestinal stromal tumors via inhibition of the BCR-ABL and c-KIT receptor tyrosine kinases, respectively
Trastuzumab	A monoclonal antibody active in ERBB2-positive breast cancers
Gefitinib	A small-molecule inhibitor of the epidermal growth factor receptor tyrosine kinase active in non-small-cell lung, hormone-refractory prostate, and head and neck cancer
Various small-molecule cyclin-dependent kinase inhibitors, e.g. flavopiridol and CYC202 (*R*-roscovitine)	Undergoing clinical evaluation
Inhibitors of RAS farnesylation, RAF-1, MEK, PI3 kinase, mTOR, and histone deacetylases	In preclinical and clinical development
Wide range of other innovative agents	In preclinical and clinical development, e.g. potential for BRAF inhibitors
17AAG	A small-molecule inhibitor of the HSP90 molecular chaperone that is completing phase I clinical with promising early results

NIH3T3 cells; various genetic and molecular studies in model organisms such as yeast, fly, and worm; and also from studies of inherited predisposition [31].

The discovery of new cancer genes should be accelerated by the impact of the Cancer Genome Project [42]. The aim here is to use the information and technologies obtained via the Human Genome Project [18, 38] to carry out a systematic, high-throughput, genome-wide screen for somatic mutations in human cancer cell lines and tissues.

The likely success of this approach is exemplified by the recent unexpected discovery that *BRAF* is an oncogene that is activated in about 70% of melanomas, 10% or more of colorectal cancers, and a smaller subset of other tumors [11]. This exciting finding, made under the auspices of the Cancer Genome Project (Sanger Centre, Hinxton, UK), indicates that the kinase encoded by the *BRAF* oncogene is an excellent target for drug discovery. One possibility is that drugs could be developed that would be selective for the mutationally activated BRAF. Such a drug would be effective in the genomically defined subset of tumors that express and are driven by the mutant kinase gene. This approach would be of particular benefit in metastatic melanoma for which therapeutic options are restricted, especially because the mutation rate is particularly high in this cancer. This discovery illustrates a number of points: (1) the power of a high-throughput genome-based approach in the discovery of new cancer genes and drug targets; (2) the potential for discovering new drugs targeted to a particular molecular pathology; (3) the value of understanding the biological function of the cancer gene and the biochemical pathway in which it operates; and (4) the downstream commercial challenges posed by the development of "niche" drug products that may have high therapeutic value but in a genomically restricted subset of cancer patients [43].

New technologies for drug discovery

Although drugs such as imatinib, trastuzumab, and gefitinib represent major technical achievements, as well as genuine medical advances, in each case there was a considerable delay between the discovery of the target and the regulatory approval of the drug. In the case of imatinib, more than 40 years elapsed between the discovery of the Philadelphia chromosome translocation and the marketing of imatinib. To accelerate drug discovery and patient benefit, the power of a range of effective, often high-throughput technologies is now being harnessed (see Figs 2 and 3).

As already discussed, high-throughput DNA sequencing and associated genomic and bioinformatic techniques are being used to speed up gene discovery and hence the identification of new molecular targets. RNAi technology is proving to be a powerful and simple means of knocking out gene function as part of target validation. Genomic and proteomic technologies are now having an impact across all areas of basic research and drug development. A particular advantage is the large number of genes, mRNAs, and proteins that can be interrogated in a single experiment. For a more extensive recent commentary on this area see Weinstein [41] and Workman [45].

High-throughput screening (HTS) is an extremely effective way of identifying small-molecule "hits" that act on a novel drug target [1]. Large compound collections from tens of thousands up to millions are required for screening campaigns involving biochemical or cell-based assays. Where the structure of the target is known or can be modeled, HTS is complemented by methods such as *in silico* screening of virtual libraries containing millions of "drug-like" compounds against the target of interest, using sophisticated computer algorithms [21]. Fragment-based screening, which involves using X-ray crystallography or nuclear magnetic resonance methods to search for very low molecular weight compounds that show weak interactions with the target, can also be profitable [5]. The use of a combination of these hit-finding methods can be highly synergistic. Following the identification of a screening hit, or more likely a series of hits against a given molecular target, the quality and potential of the hit is evaluated. Practical factors such as physicochemical properties [22], feasibility of synthesis, and overall

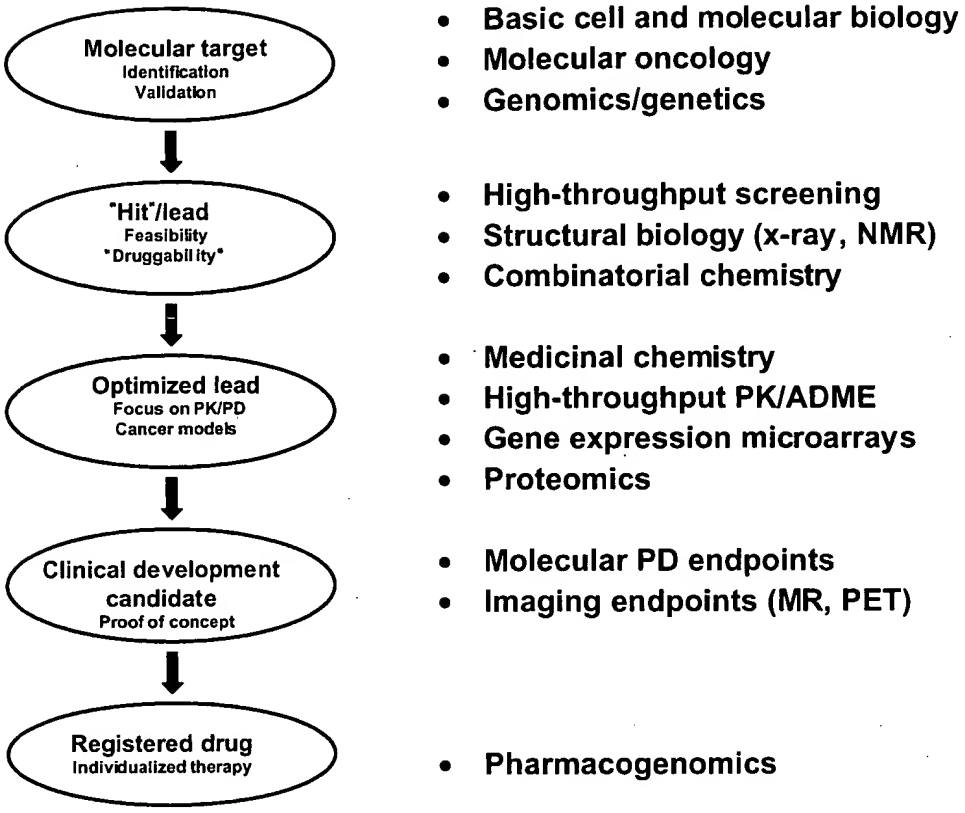


Fig. 2 The impact of new technologies at various stages of the drug discovery process (*PK* pharmacokinetics, *PD* pharmacodynamics, *NMR* nuclear magnetic resonance, *ADME* absorption, distribution, metabolism, and excretion, *MR* magnetic resonance, *PET* positron emission tomography)

"druggability" are important. Combinatorial chemistry and other new chemical methods can be used not only to create chemical diversity for HTS, but also to make more targeted libraries and for "lead explosion" to establish initial structure-activity relationships [15, 36]. Parallel synthesis methodology is valuable at this stage.

Optimization of a selected lead series towards the profile of desired properties is often focused on two main areas: (1) potency and selectivity; and (2) pharmacokinetics and absorption, distribution, metabolism, and excretion (ADME) properties. Robust assays, preferably high-throughput, need to be put in place for all these properties. These assays are formulated into a hierarchical test cascade [1]. Structure-based optimization, for example exploiting the X-ray cocrystal structure of the target–inhibitor complex, can be highly complementary to classical medicinal chemistry-based optimization. An important area for chemical innovation at the interface with bioscience is that of chemical biology [2, 37].

The ability to convert potent and selective lead compounds with activity on cancer cells in culture into agents with robust drug-like properties, particularly in terms of pharmacokinetic and metabolic properties, remains a particular challenge. It is difficult to predict such properties *ab initio*. In vitro ADME methods and higher throughput pharmacokinetic techniques, such as cassette or cocktail dosing, can be extremely valuable when used carefully with suitable lead series [33].

Mechanism of action and pharmacodynamic endpoints

It is absolutely essential during both preclinical and clinical development that particular key milestones are met. Such milestones can often constitute *go/no-go* decision points. As shown in Fig. 4, it is critical to know that active plasma and tissue concentrations of drug can be achieved in animals and patients. Next it is important to demonstrate the desired activity on the intended molecular target (e.g. kinase inhibition), followed by modulation of the corresponding biochemical pathway (e.g. RAS → ERK signaling) and also the achievement of the desired downstream biological effect (e.g. inhibition of proliferation, blockade of angiogenesis, or induction of apoptosis). Finally, these molecular and cellular events need to be linked to the therapeutic response, e.g. tumor cytostasis or regression. It is important that pharmacokinetic/pharmacodynamic relationships are established and that a pharmacological "audit trail" is constructed, consisting of measured parameters for each of the levels of analysis mentioned above (see Fig. 4, and references 46 and 48 for more details).

Pharmacodynamic endpoints may be measured on tumor biopsies or surrogate normal tissue such as peripheral blood lymphocytes, skin or buccal mucosa. Alternatively, and preferably, minimally invasive assays employing techniques such as positron emission

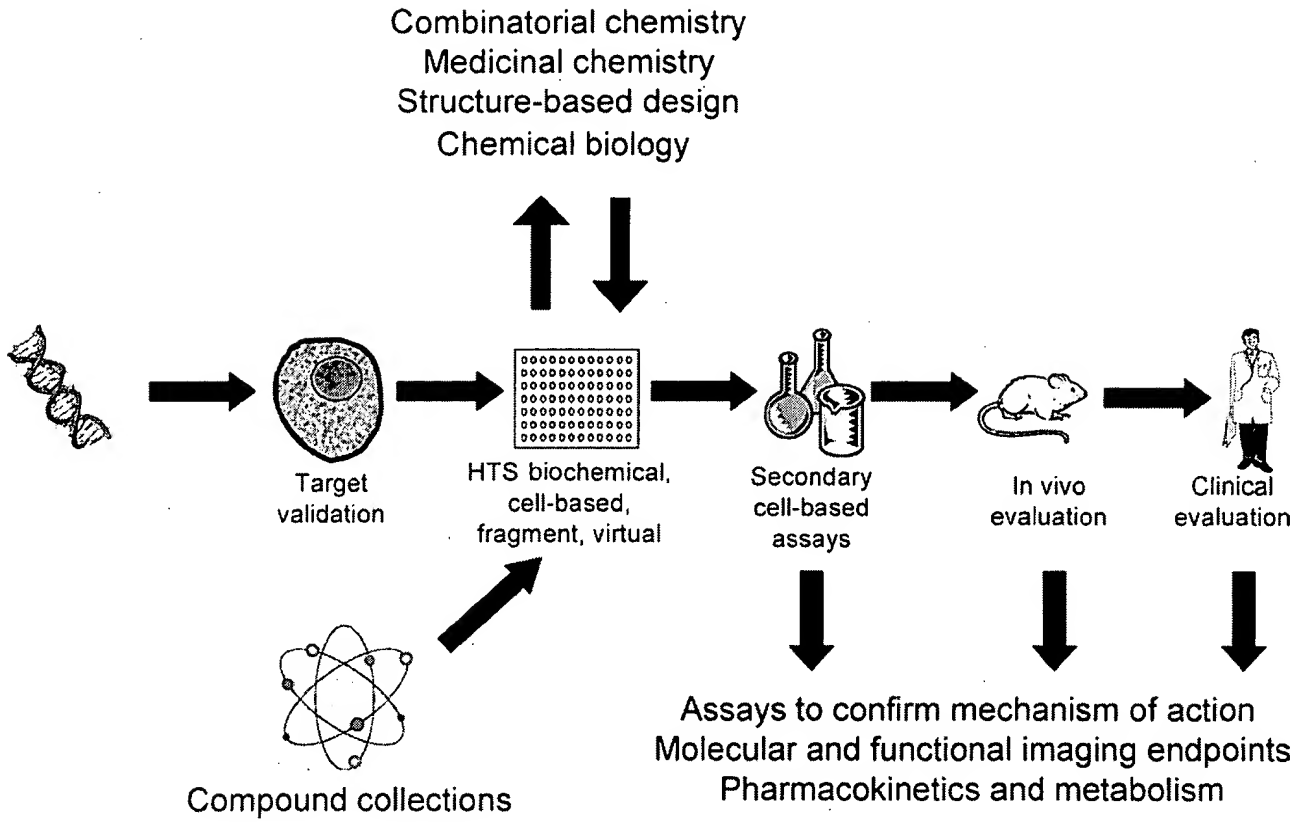


Fig. 3 Process of contemporary drug discovery (*HTS* high-throughput screening)

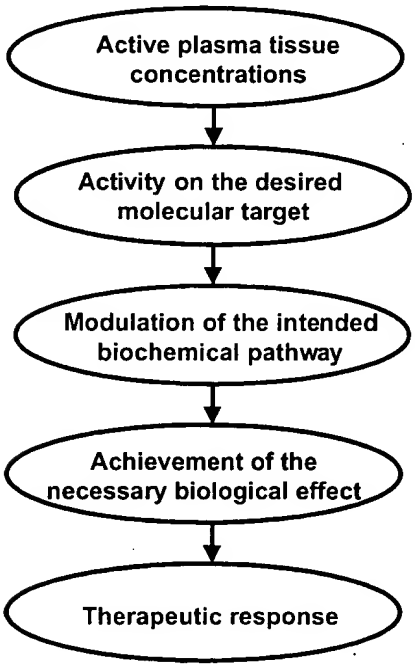


Fig. 4 Key milestones in preclinical and clinical drug development. Measurements made at each milestone allow construction of a pharmacological "audit trail" (see references 45 and 46)

tomography (PET) and magnetic resonance spectroscopy/imaging (MRS/MRI) can be extremely valuable [46, 48].

Invasive molecular endpoints can for example involve changes in protein phosphorylation, as measured by Western blotting, enzyme-linked immunosorbent assay (ELISA), or immunohistochemistry. Genome-wide expression profiling by microarray and also global proteomic analysis can provide a rich source of potential pharmacodynamic endpoints, as well as helping to understand the cellular mode of action of a drug, which may not always be as intended [8, 9, 45].

Current issues in the development of new molecular cancer therapeutics

Although rich in potential and showing signs of considerable promise, the new genome-based approach is not without its challenges (e.g. see references 3, 10, and 43). This is exemplified by the recent clinical trial results with gefitinib [14, 19]. The trials concerned were randomized, double-blind, phase III studies in which gefitinib when used in combination with chemotherapy (gemcitabine and cisplatin or paclitaxel and carboplatin) failed to improve survival in patients with chemotherapy-naive advanced non-small-cell lung cancer (NSCLC). This was perhaps surprising given that gefitinib has activity as a single agent in NSCLC, as well as in head and neck malignancy, and in hormone-refrac-

tory prostate cancer [12]. In addition, studies in preclinical models showed a benefit for the combination of gefitinib with chemotherapy. There are a number of possible explanations for the inability of gefitinib to improve clinical outcome for the particular tumor type and chemotherapy regimens concerned. One is that gefitinib and cytotoxic therapy are each maximally effective against the same tumor cell population; hence there is no additive, let alone synergistic, interaction. Another possibility is that gefitinib may block cell-cycle progression in tumor cells, thereby antagonizing the effects of cytotoxic therapy. These factors presumably outweigh potentially advantageous interactions such as blockade by gefitinib of survival pathways that might be used by cancer cells to protect themselves against cytotoxic damage. It could also be speculated that for some reason, possibly relating to changes in signaling pathways, gefitinib may be more effective in the biological context of previous exposure to chemotherapy.

Of particular potential importance is the possibility that there may be a subset of NSCLC patients who have molecular characteristics that predispose them to be responsive. This may not relate simply to the level of expression of the epidermal growth factor receptor molecular target, but could feasibly correlate with the flux through the receptor tyrosine kinase → RAS → RAF → MEK → ERK1/2 signal transduction pathway (potentially measurable using antibodies to phospho-ERK1/2) or with the expression of any number of genes that could be detected by microarray profiling. Pharmacogenomic analysis is required to identify such genes, and studies of this type will need to be an important part of the future clinical evaluation of gefitinib and other molecular therapeutics. We discuss later in this section the possibility that the optimal use of gefitinib may require a combination involving other molecular therapeutics to take out additional oncogenic pathways in NSCLC and other tumor types.

In the case of trastuzumab, although this agent clearly improves the response of ERBB2-positive breast cancer patients to cytotoxic chemotherapy, when used with anthracyclines it does have significant toxicity [12]. In addition, whereas imatinib is extremely active in the early phase of chronic myeloid leukemia (CML), it produces only short-lived responses in the accelerated and blast crisis stages of the disease; furthermore, acquired resistance to the drug is seen in chronic-phase patients, often due to mutation of the BCR-ABL kinase to a form that is no longer susceptible to imatinib [49].

One of the most important characteristics that may limit the effectiveness of signal transduction inhibitors and other molecular cancer therapeutics is the fact that the malignant progression of most cancers is probably driven by multiple oncogenic defects. Extensive epidemiological data would support the view that 5–7 rate-limiting genes are involved, although there may be as many as 10–12 oncogenic abnormalities in tumors such as pancreatic cancer. The concept of a stepwise accumulation of genetic and epigenetic abnormalities driving malignant progression is probably best exemplified in colorectal cancer [39]. Here, combinatorial oncogenesis involves a conspiracy between mutations in genes such as *RAS*, *APC*, and *P53*, which combine together to accelerate the conversion of normal cells into full-blown invasive and metastatic cancer. Although the precise source and role of genetic instability and its involvement in driving early- versus late-stage malignancy remains a highly controversial issue, there is no doubt that a high level of genetic chaos is a common feature of the major epithelial cancers such as those of the lung, breast, and bowel, as well as in the leukemias, as evidenced by the presence of large-scale amplifications, deletions, and translocations [26]. Genes involved in checkpoint control, mismatch repair, and telomere maintenance may all contribute to genomic instability and the progressive accumulation of cancer-causing defects.

The concept and reality of multistep combinatorial oncogenesis has a number of implications for the development and use of molecular cancer therapeutics. Principle among these is the issue as to whether therapeutic "correction" of a single oncogenic defect will be sufficient to achieve a significant or optimal therapeutic effect—or whether it will in fact be necessary to attend to all or at least several of the key molecular abnormalities to put the brake on combinatorial oncogenesis.

The potential problem is illustrated in Fig. 5. In the particular model example shown (Fig. 5A), the normal cell is transformed into a fully malignant cancer cell by the deregulation of three "mission-critical" pathways, most likely involving the hijacking of normal controls on proliferation signaling, cell-cycle regulation, and survival/apoptosis [13]. Pharmacological modulation of the first pathway, involving genes A–D, is without significant therapeutic effect (Fig. 5B). Similarly, intervention in the second oncogenic pathway, involving genes E–H, also confers little or no therapeutic benefit, either alone or in combination with modulation of the first pathway (Fig. 5C). However, simultaneous intervention in all three oncogenic pathways does have a major therapeutic effect (Fig. 5D). So, the model presented in Fig. 5 would predict that combinatorial oncogenesis would require combinatorial therapy. How do the data stack up against this prediction?

Surprisingly, perhaps, there are a number of published examples in which molecular correction of a single oncogenic abnormality can bring about a therapeutic effect, even in the context of multiple genetic abnormalities [40]. Examples include knockout of oncogenes such as *RAS* or *MYC*, or reintroduction of a lost tumor suppressor gene such as *P53*, *APC*, or *PTEN*. To explain such results, one can invoke the "house of cards" model and the oncogene addiction/tumor suppressor gene hypersensitivity concept [40]. In the house of cards model, the tumor requires each of the molecular abnormalities to power up malignancy; remove any one of the molecular batteries and the cancer cell collapses like a house of cards. In the related oncogene addiction/tumor suppressor gene

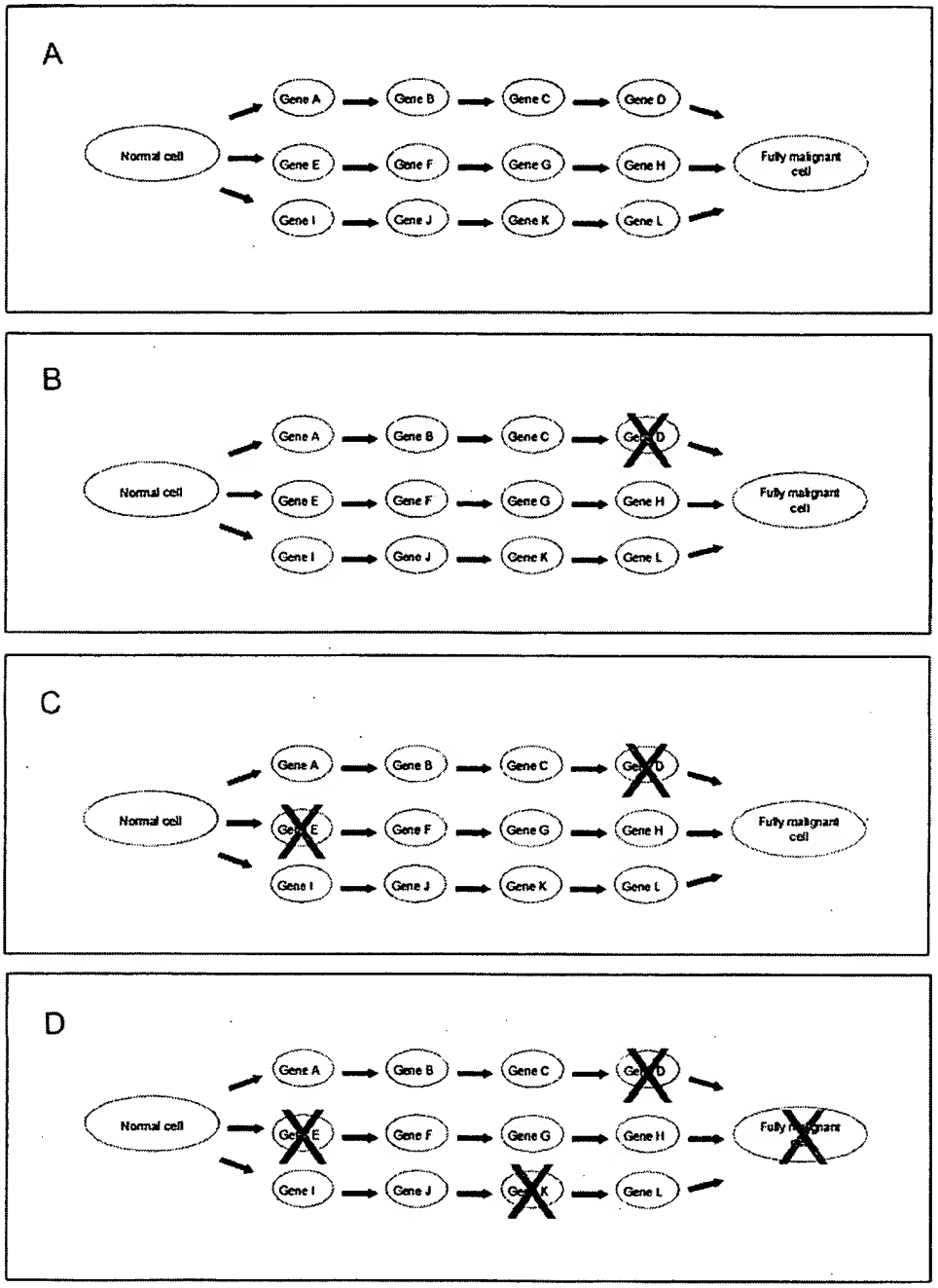


Fig. 5A–D Combinatorial oncogenesis may require combinatorial therapy. In this model, the malignancy is driven by three "mission-critical" pathways. The first pathway comprises the products of genes A–D, the second pathway the products of genes E–H, and the third pathway the products of genes I–L. As shown, the inhibition of one or two of the pathways may be insufficient for a significant therapeutic effect—combinatorial therapeutic blockade of all pathways is required for optimal treatment

hypersensitivity concept, genome instability and selection for malignancy leads to the "hard-wiring" of mission-critical oncogenic pathways and the loss of alternative or redundant signal transduction pathways. As a result, the cancer cell develops a dependence on, or addiction to, the hard-wired oncogenic pathways, together with enhanced sensitivity to reactivation of tumor suppressor functions. Because of this, treatment with a molecular therapeutic that inhibits an activated, hard-wired oncogenic pathway or reactivates a lost tumor suppressor function results in a preferential response in the cancer cell compared with its normal counterpart. It is clearly possible to invoke the oncogene addiction model to explain why a selective anticancer effect can be obtained with molecular cancer therapeutics that hit signal transduction pathways that are activated in cancer cells but that are also important for normal cell function. Probably the best example of this is the selective activity of mTOR inhibitors [e.g. rapamycin (sirolimus) derivatives] and PI3 kinase inhibitors (e.g. LY2940022) against cancer cells that have lost PTEN tumor suppressor gene function, thereby activating the PI3 kinase–AKT–mTOR pathway [27].

How does the clinical experience fit with the oncogene addiction model and the need for the correction of single versus multiple molecular abnormalities? The activity of imatinib in chronic-phase CML and gastrointestinal stromal tumors can be cited as supporting the oncogene addiction model. It is likely, however, that these are cancers in which only a single genetic defect is driving malignancy, i.e. *BCR-ABL* and mutant c-*KIT* respectively. Indeed, the lower activity in imatinib in acute and blast-phase CML and also in acute lymphocytic leukemia, where additional mutations are present, supports the view that combinations of agents may be needed to block these multiple defects. A similar argument can be made to account for the partial, although usually incomplete, responses that are seen with other molecular cancer therapeutics such as trastuzumab and gefitinib. It appears possible, then, that oncogene addiction to a single hard-wired, mission-critical pathway is partial rather than absolute. Oncogene addiction may well be present but in most cases there may be overlapping dependence on several genes and pathways. If this is correct, it would follow that treatment with a targeted drug cocktail would be advantageous. In addition, this would be likely to decrease the likelihood of resistance arising to a single agent, as seen in the clinic with imatinib in CML. This is entirely analogous to the use of multiple drug cocktails in HIV/AIDS. On the other hand, as we target several oncogenic pathways that are also used by normal cells, the key question then becomes: can we retain a therapeutic window between malignant and normal cells?

The development of HSP90 inhibitors

Given the above discussion on the likely advantage of a combinatorial blockade of multistep oncogenesis, the development of HSP90 inhibitors is brought into particularly sharp focus. The factors contributing to the "credentialing" or validation of HSP90 as a therapeutic target, together with the likely advantages of this therapeutic approach, are summarized in Table 2. HSP90 is not a product of a cancer gene per se but rather it is a protein that is required for the malignancy-driving properties of a number of bona fide oncogenes [24, 29].

The HSP90 family comprises HSP90α, HSP90β, the endoplasmic reticulum homologue GRP94, and the mitochondrial counterpart TRAP1. HSP90 is a molecular chaperone involved in protein folding. It is not, however, a generic chaperone that is required for the folding of cell proteins. Nor is it only involved under stress conditions such as heat shock. Rather, it is responsible under normal cellular conditions for the later stage folding and maintenance of the correct conformation and functional activity of a relatively restricted selection of "client" proteins. Many of the clients on this "celebrity A list" have oncogenic activity. They include several oncogenic kinases such as ERBB2, RAF-1, CDK4, POLO-1, and MET. In addition, HSP90 clients also include mutant P53, HIF-1α, estrogen/androgen receptors, and the catalytic component of telomerase hTERT. Thus inhibition of HSP90 activity leads to incorrect folding and subsequent degradation by the ubiquitin-proteasome pathway of all the above-mentioned oncogenic clients. As a result, HSP90 inhibitors are likely to block all six of the so-called "hallmark traits" of malignancy [16] and therefore have potential for one-step combinatorial therapy against a broad range of cancers. Furthermore, based on the work of Lindquist and colleagues [35], it might be speculated that inhibition of HSP90 could uncover synthetic lethal mutations in cancer cells.

Encouragingly for the approach, certain natural products that were known to have anticancer activity were found to target HSP90 [24, 29]. In particular, these include radicicol and geldanamycin (see Fig. 6 for chemical structures). These agents work by competing

Table 2 HSP90 target validation (for further details see reference 24)

Molecular chaperone involved in protein folding
Overexpressed in human tumors (e.g. due to stress and oncoproteins)
Essential for stability and function of many oncogenic "client" proteins e.g. ERBB2, RAF-1, CDK4, POLO-1, MET, mutant P53, HIF1α, estrogen/androgen receptors, and telomerase hTERT
Inhibition likely to block all six "hallmark traits" of cancer
Potential for one-step combinatorial therapy against a broad range of malignancies
May uncover synthetic lethal mutations in cancers
Natural products that target HSP90 have anticancer activity
Proof of concept for therapeutic selectivity demonstrated in human tumor xenograft models
First-in-class inhibitor 17AAG now showing evidence of biological and clinical activity at well-tolerated doses

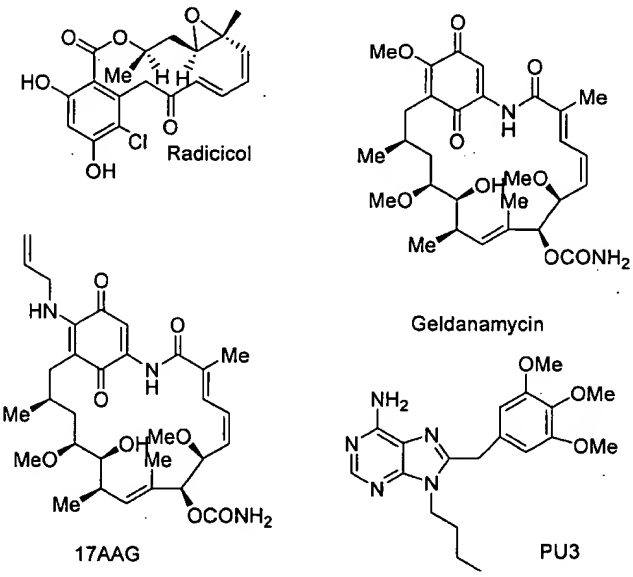


Fig. 6 Chemical structures of HSP90 inhibitors

with ATP for binding at the nucleotide-docking site located in the N-terminal domain of HSP90 [32, 34]. ATP binding and hydrolysis are essential for the functioning of the chaperone and drug binding prevents the correct assembly of mature HSP90/client protein/cochaperone complexes. This appears to result in recruitment of a ubiquitin ligase to the immature complex, leading to proteasomal degradation of client protein [24].

Proof of concept for therapeutic selectivity towards cancer cells was exemplified with the geldanamycin analog 17AAG (Fig. 6) in human tumor xenograft models grown in immunosuppressed mice [20]. Furthermore, 17AAG has entered clinical trials as the first-in-class inhibitor of HSP90 and is now showing consistent molecular evidence of the desired mechanism of action, together with early indications of therapeutic activity [4].

We have shown that treatment of human colon cancer cells with 17AAG leads to combinatorial depletion of key oncogenic client proteins such as RAF-1 and AKT, consistent with the demonstrated inhibition of the ERK1/2 and PI3 kinase signaling pathway and the downstream induction of cell-cycle arrest and apoptosis [8, 17].

We have used global gene expression microarray profiling to investigate genes that might be involved in sensitivity to 17AAG, as well as to identify potential pharmacodynamic markers of effective HSP90 inhibition [8]. In addition, we used proteomic analysis to identify global responses to HSP90 inhibition by 17AAG at the protein level (collaboration with Professor Mike Waterfield and colleagues, Ludwig Institute for Cancer Research, University College London, London, UK). A molecular signature of HSP90 inhibition has been defined, consisting of depletion of client proteins such as RAF-1, CDK4, and ERBB2 at the protein level (with no effect at the mRNA level) together with upregulation of HSP70 at both the mRNA and protein levels [24]. In some cancer cell lines, HSP90 itself is upregulated. We routinely determine the molecular signature of HSP90 inhibition by Western blotting. In addition, we are also developing ELISA assays for greater sensitivity and more straightforward quantification.

In terms of the expression of genes that may confer sensitivity or resistance, we have shown that high levels of the quinone reductase NQO1/DT-diaphorase cause considerable sensitization toward 17AAG, which has a 17-allylamino group, although not to the major metabolite of 17AAG, which has an amino moiety at the 17 position, or to geldanamycin, which has a methoxy group at the 17 position [20]. The results suggest a role for activation via quinone metabolism, although the HSP90 mechanism is retained. Further work is required to elucidate the details and full significance of the effect.

Interestingly, our studies have also suggested that tumor lines that respond to treatment by expressing increased levels of the HSP90 target itself may recover more rapidly from the effects of 17AAG and therefore be less sensitive to the drug [8].

In collaborative studies published recently, we have identified the new gene product AHA1 as a novel cochaperone that activates the essential ATPase activity of HSP90 and which is upregulated in human tumor cells by stress, heat shock, and pharmacological HSP90 inhibitors [30]. Using a combination of gene expression microarrays, proteomics (two-dimensional gel electrophoresis with MALDI mass spectrometry) and Western blotting, we showed that *AHA1* gene expression is upregulated at the level of both mRNA and protein in response to treatment of human tumor cells with the HSP90 inhibitors radicicol and 17AAG. The mechanistic, pharmacological, and therapeutic significance of these observations is now under investigation.

Having shown good activity in xenograft models and an acceptable therapeutic index in animal models, 17AAG has been taken into clinical trials in our own institution and at our four centers in the USA under the auspices of the US National Cancer Institute and Cancer Research UK (formerly the Cancer Research Campaign). In the UK trial at the Cancer Research UK Centre for Cancer Therapeutics, Institute of Cancer Research, and the Royal Marsden Hospital [4], 17AAG has been given weekly by intravenous infusion at doses up to 450 mg/m^2/week. Pharmacokinetic studies show that plasma concentrations are above the IC_{50} for inhibition of tumor cell growth for prolonged periods. In addition, depletion of RAF-1, CDK4, and the SRC family kinase LCK has been clearly demonstrated in peripheral blood lymphocytes, together with upregulation of HSP70. Furthermore, depletion of RAF-1 and CDK4 alongside increased expression of HSP70 has also been observed in malignant tissue by comparing tumor biopsies taken before and after treatment. Consistent with these molecular changes, we have seen evidence of disease stabilization in some patients. RNA has been prepared from certain tumor biopsies to allow global expression profiling to be carried out. This should generate valuable results to compare with those from in vitro cell-culture exposures [8].

Although relatively invasive assays are providing valuable information by demonstrating that 17AAG is able to inhibit its molecular target both in peripheral blood lymphocytes and in tumor biopsy material, minimally invasive assays such as those involving PET and MRS/MRI would have major advantages [46, 48]. In collaboration with Professors Martin Leach, John Griffiths, and colleagues (Cancer Research UK Biomedical Magnetic Resonance Group, St George's Hospital Medical School, London, and Cancer Research UK Clinical Magnetic Resonance Research Group, Institute of Cancer Research and Royal Marsden Hospital, Sutton, UK), we have noted interesting changes in human xenograft tumors following treatment with 17AAG, in particular an unusual increase in the levels of phosphoethanolamine and phosphocholine [7]. These may be indicative of alterations in lipid signaling and/or membrane turnover. In addition, we are collaborating with Professor Pat Price and Dr. Eric Aboagye (Cancer Research UK PET Oncology Group, Molecular Imaging

Centre, Manchester, and Cancer Research UK PET Oncology Group, MRC Cyclotron Unit, Hammersmith Hospital, Imperial College School of Medicine, London, UK) to use labeled choline PET tracers to monitor the effects of 17AAG in tumors [23]. Overall, the potential to use molecular or functional imaging to monitor the pharmacodynamic effects of the new molecular cancer therapeutics is an exciting area.

17AAG shows significant promise and demonstrates proof of concept for HSP90 inhibition in humans. It does, however, have a number of potential limitations. These include:

- Limited stability and complex formulation
- Modest potency against the HSP90 target
- Substrate for P-glycoprotein
- Activated by polymorphic NQO1/DT-diaphorase
- Metabolism by polymorphic cytochrome P450
- Low oral bioavailability
- Limited therapeutic index

Because of these potential issues, several groups are seeking small-molecule, synthetic inhibitors of HSP90 as alternatives to the existing natural products. A range of approaches are likely to be taken, including those described earlier in this commentary and depicted in Fig. 3.

One interesting lead that has emerged is the synthetic purine-based compound PU3 (Fig. 6). This agent has been shown to inhibit HSP90 in cancer cells and to retard their growth [6]. PU3 appears to behave like the natural product agents, competing with ATP at the nucleotide-binding site of the N-terminal domain of HSP90 [6]. Another interesting compound is novobiocin. This appears to act in a different way by binding to the C-terminal domain of HSP90 [25]. Given the attractiveness of the target and the encouraging results with 17AAG, it appears likely that more synthetic chemical inhibitors of HSP90 will emerge.

There are many challenges ahead with HSP90 inhibitors. Some of the important outstanding questions include:

- What is the optimal treatment regimen?
- How should the drug be used as a single agent?
- How should the drug be used in combination with cytotoxics, e.g. paclitaxel [28]?
- Will any tumor types be particularly sensitive?
- Are any particular client proteins especially important for response in certain tumor settings?
- Will particular genomic abnormalities predispose to sensitivity or resistance?

Conclusions

The following overall conclusions can be drawn:

- Proof of principle is now established that targeting cancer genome abnormalities and the molecular pathology of cancer can be clinically beneficial.
- New molecular targets continue to emerge from cancer genomics.
- Blocking multistep oncogenesis will most likely require combinatorial therapies.
- This may be delivered in individualized cocktails of molecularly targeted agents.
- HSP90 inhibitors such as 17AAG may block multiple oncogenic pathways in a single drug.
- Deployment of multidisciplinary skills and new technologies is required to accelerate the pace and improve the efficiency of drug discovery against new molecular targets.
- Clinical development strategies must pay close attention to the proposed mechanism of action and a pharmacological audit trail must be constructed to allow rational decision-making, including *go/no-go*.
- Demonstration of proof of concept is invaluable in hypothesis testing phase I clinical trials.
- Pharmacodynamic and pharmacogenomic markers are essential for success.

The explosion of new molecular targets and the development and application of many powerful technologies should accelerate the discovery of innovative molecular therapeutics. There are many challenges ahead and the risks associated with each individual agent remain considerable, but the prospects for overall success with individualized therapies targeted to the molecular pathology of the individual patient are excellent [47, 49]. This exciting translational work requires many disciplines (e.g. chemistry, biology, and medicine) and organizations (e.g. academia, biotech, and large pharmaceutical companies) to work together internationally to accelerate patient benefit.

Acknowledgements The work of the author and the Centre for Cancer Therapeutics, Institute of Cancer Research (http://www.icr.ac.uk/cctherap/index.html) is funded by a core grant (C309/A2984) from Cancer Research UK (http://www.cancerresearchuk.org) and the author is a Cancer Research UK Life Fellow. I thank my colleagues and coworkers for their collaboration and stimulating discussions. I also thank Dr. Ted McDonald and colleagues in the Cancer Research UK Centre for Cancer Therapeutics for Fig. 6.

References

1. Aherne GW, McDonald E, Workman P (2002) Finding the needle in the haystack: why high-throughput screening is good for your health. Breast Cancer Res 4:148
2. Alaimo PJ, Shogren-Knaak MA, Shokat KM (2001) Chemical genetic approaches for the elucidation of signaling pathways. Curr Opin Chem Biol 5:360
3. Atkins JH, Gershell LJ (2002) Selective anticancer drugs. Nat Rev Drug Discov 1:491
4. Banerji U, O'Donnell A, Scurr M, Benson C, Hanwell J, Clark S, Raynaud F, Turner A, Walton M, Workman P, Judson I (2001) Phase I trial of the heat shock protein 90 (HSP90) inhibitor 17-allylamino-17-demethoxygeldanamycin (17AAG). Pharmacokinetic (PK) profile and pharmacodynamic (PD) endpoints. Proc Am Soc Clin Oncol Abstract 326

5. Carr R, Jhoti H (2002) Structure-based screening of low-affinity compounds. Drug Discov Today 7:522
6. Chiosis G, Timaul MN, Lucas B, Munster PN, Zheng FF, Sepp-Lorenzino L, Rosen N (2001) A small molecule designed to bind to the adenine nucleotide pocket of Hsp90 causes Her2 degradation and the growth arrest and differentiation of breast cancer cells. Chem Biol 8:289
7. Chung Y-L, Troy H, Banerji U, Judson IR, Leach MO, Stubbs M, Ronen S, Workman P, Griffiths JR (2002) The pharmacodynamic effects of 17-AAG on HT29 xenografts in mice monitored by magnetic resonance spectroscopy. Proc Am Assoc Cancer Res Abstract 371
8. Clarke PA, Hostein I, Banerji U, Di Stefano F, Maloney A, Walton M, Judson I, Workman P (2000) Gene expression profiling of human colon cancer cells following inhibition of signal transduction by 17-allylamino-17-demethoxygeldanamycin, an inhibitor of the hsp90 molecular chaperone. Oncogene 19:4125
9. Clarke PA, te Poele R, Wooster R, Workman P (2001) Gene expression microarray analysis in cancer biology, pharmacology, and drug development: progress and potential. Biochem Pharmacol 62:1311
10. Couzin J (2002) Cancer drugs: smart weapons prove tough to design. Science 298:522
11. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, Davis N, Dicks E, Ewing R, Floyd Y, Gray K, Hall S, Hawes R, Hughes J, Kosmidou V, Menzies A, Mould C, Parker A, Stevens C, Watt S, Hooper S, Wilson R, Jayatilake H, Gusterson BA, Cooper C, Shipley J, Hargrave D, Pritchard-Jones K, Maitland N, Chenevix-Trench G, Riggins GJ, Bigner DD, Palmieri G, Cossu A, Flanagan A, Nicholson A, Ho JW, Leung SY, Yuen ST, Weber BL, Seigler HF, Darrow TL, Paterson H, Marais R, Marshall CJ, Wooster R, Stratton MR, Futreal PA (2002) Mutations of the *BRAF* gene in human cancer. Nature 417:949
12. de Bono JS, Rowinsky EK (2002) The ErbB receptor family: a therapeutic target for cancer. Trends Mol Med [Suppl] 8:S19
13. Evan GI, Vousden KH (2001) Proliferation, cell cycle and apoptosis in cancer. Nature 411:342
14. Giaconne G, Johnson DH, Manegold C, Scagliotti GV, Rosell R, Wolf M, Rennie P, Ochs J, Averbuch S, Fandi A (2002) A phase III clinical trial of ZD1839 ('Iressa') in combination with gemcitabine and cisplatin in chemotherapy-naive patients with advanced non-small-cell lung cancer (INTACT 1) (abstract 4O). Ann Oncol 13 [Suppl 5]:2
15. Gullier F, Orain D, Bradley M (2000) Linkers and cleavage strategies in solid-phase organic synthesis and combinatorial chemistry. Chem Rev 100:2091
16. Hanahan D, Weinberg RA (2000) The hallmarks of cancer. Cell 100:57
17. Hostein I, Robertson D, Di Stefano F, Workman P, Clarke PA (2001) Inhibition of signal transduction by the Hsp90 inhibitor 17-allylamino-17-demethoxygeldanamycin results in cytostasis and apoptosis. Cancer Res 61:4003
18. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860
19. Johnson DH, Herbst R, Giaconne G, Schiller J, Natale RB, Miller V, Wolf M, Helton A, Averbuch S, Grous J (2002) ZD1839 ('Iressa') in combination with paclitaxel and carboplatin in chemotherapy-naive patients with advanced non-small-cell lung cancer (NSCLC): results from a phase III clinical trial (INTACT 2) (abstract 468O). Ann Oncol 13 [Suppl 5]:127
20. Kelland LR, Sharp SY, Rogers PM, Myers TG, Workman P (1999) DT-Diaphorase expression and tumor cell sensitivity to 17-allylamino-17-demethoxygeldanamycin, an inhibitor of heat shock protein 90. J Natl Cancer Inst 91:1940
21. Leach AR, Hann MM (2000) The *in silico* world of virtual libraries. Drug Discov Today 5:326
22. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 23:3
23. Liu D, Hutchinson OC, Osman S, Price P, Workman P, Aboagye EO (2002) Use of radiolabelled choline as a pharmacodynamic marker for the signal transduction inhibitor geldanamycin. Br J Cancer 87:783
24. Maloney A, Workman P (2002) HSP90 as a new therapeutic target for cancer therapy: the story unfolds. Expert Opin Biol Ther 2:3
25. Marcu MG, Chadli A, Bouhouche I, Catelli M, Neckers LM (2000) The heat shock protein 90 antagonist novobiocin interacts with a previously unrecognized ATP-binding domain in the carboxyl terminus of the chaperone. J Biol Chem 275:37181
26. Marx J (2002) Debate surges over the origins of genomic defects in cancer. Science 297:544
27. Mills GB, Lu Y, Kohn EC (2001) Linking molecular therapeutics to molecular diagnostics: inhibition of the FRAP/RAFT/TOR component of the PI3 K pathway preferentially blocks PTEN mutant cells in vitro and in vivo. Proc Natl Acad Sci U S A 98:10031
28. Munster PN, Basso A, Solit D, Norton L, Rosen N (2001) Modulation of Hsp90 function by ansamycins sensitizes breast cancer cells to chemotherapy-induced apoptosis in an RB- and schedule-dependent manner. Clin Cancer Res 7:2228
29. Neckers L (2002) Hsp90 inhibitors as novel cancer chemotherapeutic agents. Trends Mol Med 8:S55
30. Panaretou B, Siligardi G, Meyer P, Maloney A, Sullivan JK, Singh S, Millson SH, Clarke PA, Naaby-Hansen S, Stein R, Cramer R, Mollapour M, Workman P, Piper PW, Pearl LH, Prodromou C (2002) Activation of the ATPase activity of Hsp90 by the stress-regulated cochaperone Aha1. Mol Cell 10:1307
31. Ponder BA (2001) Cancer genetics. Nature 411:336
32. Prodromou C, Roe SM, O'Brien R, Ladbury JE, Piper PW, Pearl LH (1997) Identification and structural characterization of the ATP/ADP-binding site in the Hsp90 molecular chaperone. Cell 90:65
33. Raynaud FI, Goddard P, Fischer P, McClue S, Workman P (2001) Cassette dosing pharmacokinetics for 107 analogues of the trisubstituted CDK2 inhibitor roscovitine. Proc Am Assoc Cancer Res Abstract 2056
34. Roe SM, Prodromou C, O'Brien R, Ladbury JE, Piper PW, Pearl LH (1999) Structural basis for inhibition of the Hsp90 molecular chaperone by the antitumor antibiotics radicicol and geldanamycin. J Med Chem 42:260
35. Rutherford SL, Lindquist S (1998) Hsp90 as a capacitor for morphological evolution. Nature 396:336
36. Schreiber SL (2000) Target-oriented and diversity-oriented organic synthesis in drug discovery. Science 287:1964
37. Stockwell BR, Haggarty SJ, Schreiber SL (1999) High-throughput screening of small molecules in miniaturized mammalian cell-based assays involving post-translational modifications. Chem Biol 6:71
38. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, et al (2001) The sequence of the human genome. Science 291:1304
39. Vogelstein B, Kinzler K (1993). The multistep nature of cancer. Trends Genet 9:138
40. Weinstein IB (2002) Cancer: addiction to oncogenes—the Achilles heal of cancer. Science 297:63
41. Weinstein JN (2002) 'Omic' and hypothesis-driven research in the molecular pharmacology of cancer. Curr Opin Pharmacol 2:361
42. Wooster R (2001) Richard Wooster on cancer and the Human Genome Project. Lancet Oncol 2:176
43. Workman P (2000) Cancer—21st century solutions. Biotech Invest Today 1:28
44. Workman P (2001) Scoring a bull's-eye against cancer genome targets. Curr Opin Pharmacol 1:342

45. Workman P (2002) The impact of genomic and proteomic technologies on the development of new cancer drugs. Ann Oncol 13 [Suppl 4]:115
46. Workman P (2002) Challenges of PK/PD measurements in modern drug development. Eur J Cancer 38:2189
47. Workman P (2002) Cancer genome targets: RAF-ing up tumour cells to overcome oncogene addiction. Expert Rev Anticancer Ther 2:611
48. Workman P (2003) How much gets there and what does it do? The need for better pharmacokinetic and pharmacodynamic endpoints in contemporary drug discovery and development. Curr Pharm Des (in press)
49. Workman P, Kaye SB (eds) (2002) A Trends Guide to Cancer Therapeutics. Trends Mol Med [Suppl] 8

Mellerson, Kendra

From: Gakh, Yelena
Sent: Tuesday, August 05, 2003 2:33 PM
To: STIC-EIC1700
Subject: 09890973

Dear Kendra:

please order one more list:

2. TITLE: "Metabolomic analysis of the consequences of cadmium exposure in Silene cucubalus cell cultures via 1H NMR spectroscopy and chemometrics"
AUTHOR(S): *Bailey, Nigel J. C.; Oven, Matjaz; Holmes, Elaine; Nicholson, Jeremy K.; Zenk, Meinhart H.*
CORPORATE SOURCE: Technology and Medicine, Imperial College of Science, Biomedical Sciences Division, Biological Chemistry, University of London, London, SW7 2AZ, UK
SOURCE: **Phytochemistry (Elsevier) (2003), 62(6), 851-858**

Thank you,

Yelena

Biotech
QK861.P45

Yelena G. Gakh, Ph.D.

Patent Examiner
USPTO, cp3/7B-08
(703)306-5906

Available online at www.sciencedirect.com
SCIENCE DIRECT®
PHYTOCHEMISTRY
Phytochemistry 62 (2003) 851–858
www.elsevier.com/locate/phytochem

Metabolomic analysis of the consequences of cadmium exposure in *Silene cucubalus* cell cultures via ^{1}H NMR spectroscopy and chemometrics

Nigel J.C. Bailey[a,*], Matjaz Oven[b,1], Elaine Holmes[a], Jeremy K. Nicholson[a], Meinhart H. Zenk[c]

[a] *Biological Chemistry, Biomedical Sciences Division, Imperial College of Science, Technology and Medicine, University of London, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ, UK*
[b] *Leibniz-Institut für Pflanzenbiochemie, Weinberg 3, D-06120 Halle/Saale, Germany*
[c] *Biozentrum, Pharmazie, Universität Halle, Weinbergweg 22, D-06120 Halle/Saale, Germany*

Received 5 June 2002; received in revised form 5 August 2002

Abstract

Several essential and non-essential metals (typically those from periods 4, 5 and 6 in groups 11–15 in the periodic table) are commonly detoxified in higher plants by complexation with phytochelatin. The genetic and gross metabolic basis of metal tolerance in plants is, however, poorly understood. Here, we have analyzed plant cell extracts using ^{1}H NMR spectroscopy combined with multivariate statistical analysis of the data to investigate the biochemical consequences of Cd^{2+} exposure in *Silene cucubalus* cell cultures. Principal components analysis of ^{1}H NMR spectra showed clear discrimination between control and Cd^{2+} dosed groups, demonstrating the metabolic effects of Cd^{2+} and thus allowing the identification of increases in malic acid and acetate, and decreases in glutamine and branched chain amino acids as consequences of Cd^{2+} exposure. This work shows the value of NMR-based metabolomic approaches to the determination of biochemical effects of pollutants in naturally selected populations.

Keywords: Cadmium; Metabolomics; NMR spectroscopy; *Silene cucubalus*; Metabolite

1. Introduction

The development of novel analytical strategies for deriving information on differential gene function in relation to environmental stressors is essential in order to advance the molecular basis of metal tolerance. Whereas genomics and proteomics can provide insights into the potential of a biological system to interact with external perturbations (pharmaceutical/agrochemical compounds, pollutants, environmental effects), it is the resulting changes in the metabolic profile of the system that are potentially more use for the understanding of the biochemical reaction to stress. This is because it is changes in the metabolic profile that are the ultimate result of such external influences. '*Metabonomics*', defined as "the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification" (Nicholson et al., 1999, 2002; Lindon et al., 2001), is increasingly being used for the analysis of a range of biological problems including toxicological assessment (Holmes et al., 2001), differentiation between genetic strains (Gavaghan et al., 2000), comparative mammalian biochemistry (Griffin et al., 2000) and natural product characterization (Bailey et al., 2002; Belton et al., 1998). In parallel, there have been developments in '*Metabolomics*', which broadly encompasses the study of the metabolic response in isolated systems as opposed to the whole system approach

* Corresponding author at: SCYNEXIS Europe Limited, Fyfield Business & Research Park, Fyfield Road, Ongar, Essex CM5 0GS, UK. Tel.: +44-1277-367036.
E-mail address: nigel.bailey@scynexis.com, http://www.med.ic.ac.uk/divisions/1/home.asp (N.J.C. Bailey).
URL: http://www.med.ic.ac.uk/divisions/1/home.asp
[1] Present address: School of Biological Sciences, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK.

0031-9422/03/$ - see front matter
doi:10.1016/S0031-9422(02)00719-7

described by metabonomics. Metabolomic studies have been reported on the analysis of the consequences of genetic manipulation and strain differentiation at the cellular level, for example in the characterization of phenotypic differences in strains of yeast (Raamsdonk et al., 2001). While the application of NMR spectroscopy to metabonomic investigations has gained momentum, relatively little data have hitherto been published on the application of high resolution NMR spectroscopy in plant metabolomics. It has been reported, however, that a combination of off-line HPLC–NMR spectroscopy with rudimentary data analysis has been employed for the evaluation of metabolic changes in transgenic food crops (Noteborn et al., 2000). Several recent studies have shown the application of metabolomic-type analyses using GC–MS for the analysis of transgenic potato tubers (Roessner et al., 2000, 2001) and *Arabidopsis* genotypes (Fiehn et al., 2000). While MS-based detection techniques typically display greater analytical sensitivity than NMR spectroscopic detection, there is an inherent necessity for the analyte of interest to ionize in the mass spectrometer along with requirements for pre-analysis derivatization. This means that the non-selective, yet highly specific approach of NMR spectroscopy, where no pre-judgement of the sample is required, offers several advantages with respect to the development of an analytical methodology that is readily transferable between samples from differing applications. Here we demonstrate the value of NMR based metabolomics in the investigation of metal tolerance and toxicity in plants, specifically, the effects of cadmium on *Silene cucubalus*.

Cadmium is a putatively non-essential and potentially highly toxic element to all classes of living organisms. Soils and water may be contaminated with Cd^{2+} as a result of mining or industrial activities, use of phosphorus containing fertilizers, land applications of sewage sludge, and atmospheric deposition (di Toppi and Gabbrielli, 1999). Soil contamination of Cd^{2+} presents a significant concern as increased Cd^{2+} bioavailability may harm ecosystem functions, or result in an unacceptable level of transfer of Cd^{2+} to the food chain. Cadmium exposure results in lesions in the kidneys of higher vertebrates and man (Nicholson et al., 1983; Nicholson and Osborn, 1983). Recent research (Lombi et al., 2000) has shown that several plant species may be Cd-tolerant and indeed, one plant species (*Thlaspi caerulescens*, a Brassicaceae) has been identified as being a Cd^{2+} hyperaccumulator (defined as storing >100 mg Cd^{2+} kg^{-1} in the shoot dry matter). *S. cucubalus* is known to respond to cadmium exposure through the chelation of metal ions by a family of peptide ligands, the phytochelatins, which consist of repetitions of γ-Glu-Cys sequences with a terminal Gly (Grill et al., 1985; Zenk, 1996; Cobbett, 2000). However, despite the evidence for phytochelatin involvement, little is known about the gross changes in biochemical status in *S. cucubalus* cultures as a result of Cd^{2+} exposure. The aim of this work was to apply an NMR-based metabolomic approach to investigate the metabolic responses of *S. cucubalus* following Cd^{2+} exposure in vitro.

2. Results and discussion

2.1. ^1H NMR spectroscopic analysis of the samples

The ^1H NMR spectra for the predose (samples obtained on day 0, at the time of transfer into fresh media), control (samples obtained on day 3 at same time as dosed samples were obtained) and dosed (samples obtained on day 3 following exposure to 150 μM Cd^{2+}) are shown in Figs. 1a–c, respectively. It was possible to observe clear differences between these spectra, indicating changes in biochemical status with respect to time, i.e. between predose (a) and control (b) samples, where there is a time difference of three days and following exposure to the cadmium i.e. between control (b) and dosed (c) samples. Although differences between the spectra were readily observed, it was important to derive metabolic differences between sample classes based on the mathematical variance in the matrix rather than solely through visual inspection, hence the use of principal components analysis (PCA) to reduce the dimensionality of the data thus allowing easier interpretation of the results.

2.2. Pattern recognition analysis of the ^1H NMR spectra

PCA is an unsupervized method, i.e. analysis is performed without use of knowledge of sample class, which reduces the dimensionality of the data input whilst expressing much of the original *n*-dimensional variance in a 2- or 3-D map (Eriksson et al., 1999). By producing new linear combinations of the original variables, i.e. the integrated NMR spectral regions, it is possible to plot such data in order to indicate relationships between samples in the multidimensional space. The result is a diagram known as a scores plot that can be used to determine the similarities and differences between many samples Fig. 2). This dataset of NMR spectra from the cell culture extracts displayed good discrimination between the three classes analyzed, in that the classes were easily differentiated from one another. Further this separation took place in the first two principal components (PCs[2]) which cumulatively account for 96.5% of the variance in the dataset, indicating that it

[2] The abbreviation PC is in common usage to refer to both principal components and phytochelatins. PC is used to refer to principal components only throughout this work.

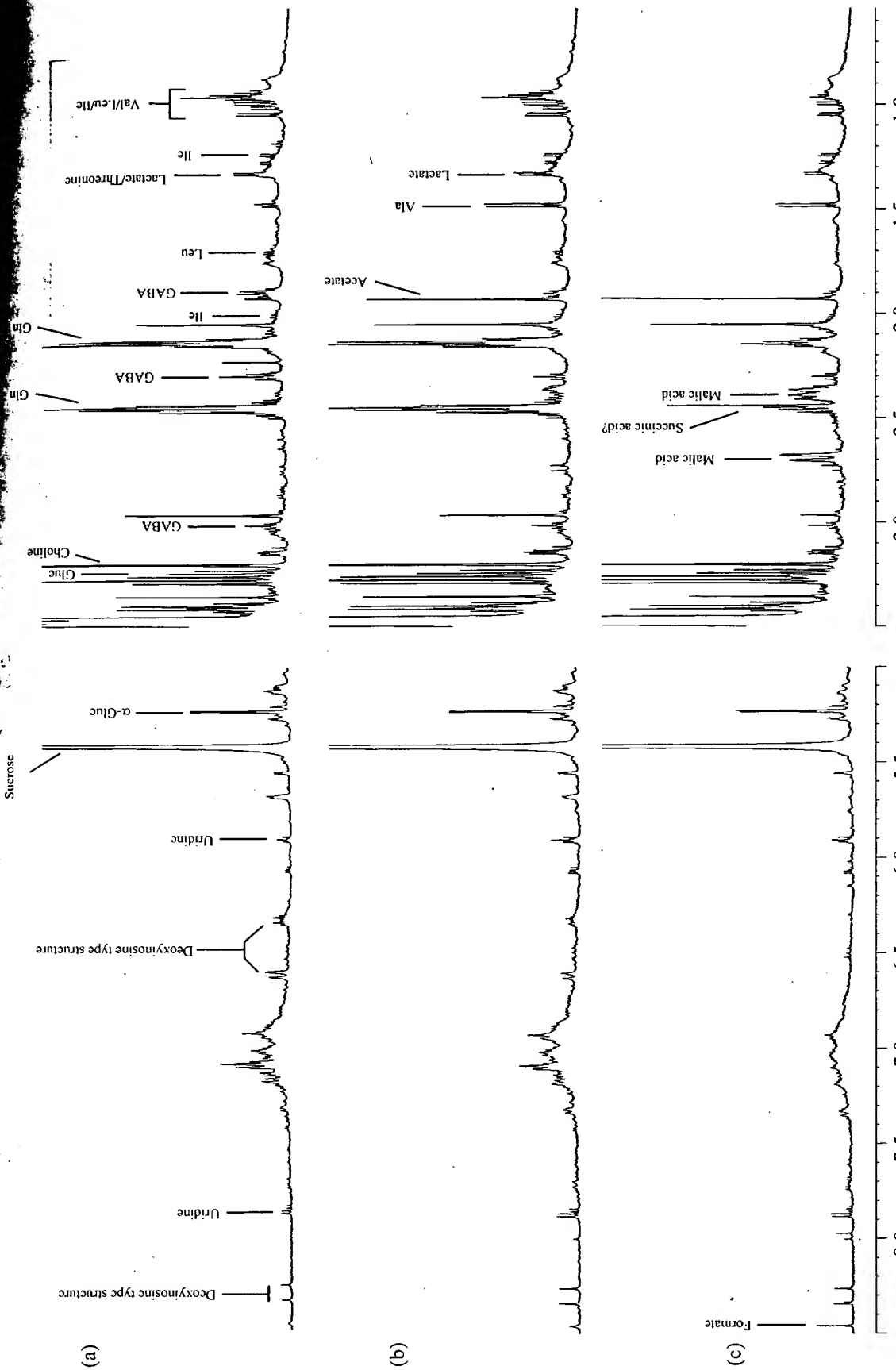


Fig. 1. NMR spectra for all three classes (a) predose (day 0 after new growing media was added), (b) control (day 3), (c) dosed (day 3 after addition of 150 μM Cd^{2+}). Region containing residual HOD and sucrose resonances (present in the media solution) are removed for clarity.

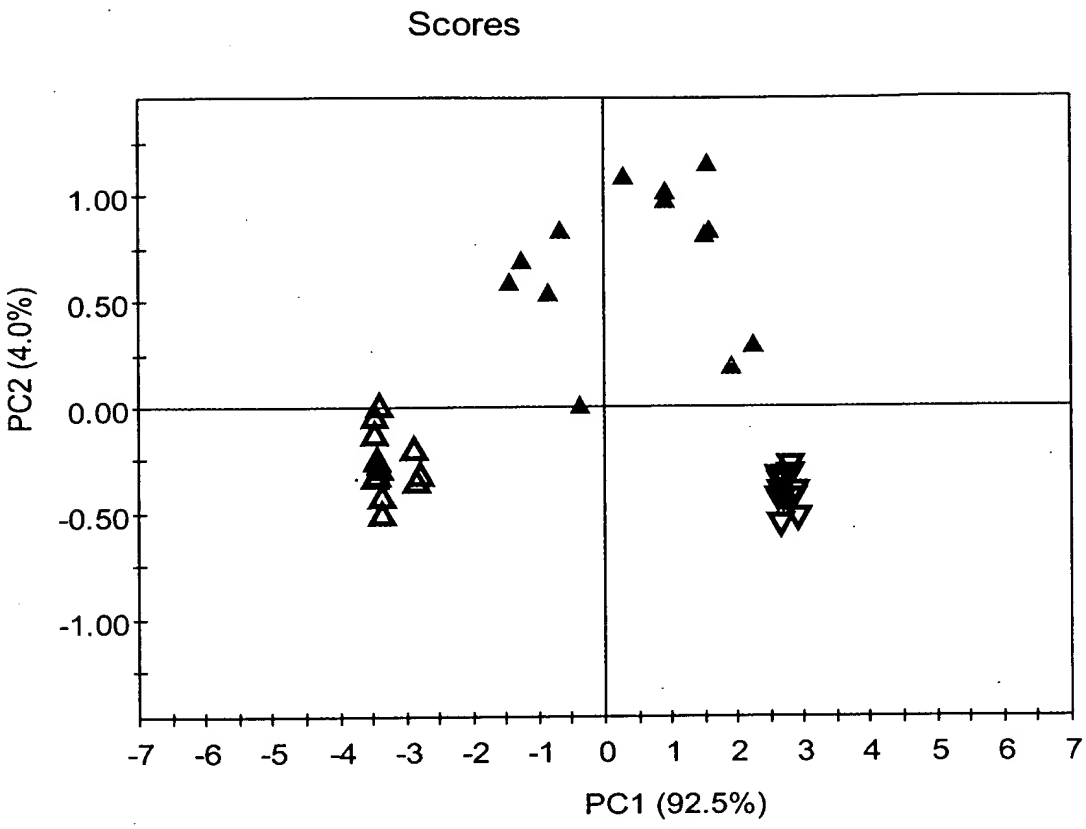


Fig. 2. Scores plot (PC1 v PC2) for predose (open red triangles), control (closed blue triangles) and dosed (open purple triangles) sample groups following PC analysis. The plot displays clear discrimination between the three groups, accounting for nearly 97% of the variance within the dataset.

the difference between the three classes analyzed that is the major discriminating factor between samples rather than any other unrelated variation between samples. There are differences in metabolic profile due to both dosing and incubation time in the absence of Cd^{2+}. The time-related changes reflect adaptation to the new growth/nutrient conditions in the culture flasks. Both the predose and dosed sample classes were tightly grouped together within their classes (Figs. 2 and 3), whereas the control data were much more diffuse (standard deviations for predose and dosed samples in PC1 were 0.1 and 0.2 respectively, while for control samples it was 1.3. For PC2, the values were 0.1, 0.2 and 0.4 respectively). It can be seen that at the start of the study the samples in the predose class are biochemically similar to each other (relative to the samples in the control group). After 3 days of growth the controls separate from the predose condition and the samples have also biochemically diverged with respect to each other, resulting in the larger standard deviations indicated above. The effects of the Cd^{2+}-exposure on the cellular metabolic profiles were markedly larger than the differences caused by the 'natural' divergence of the control and predose groups. The Cd^{2+} dosed group formed a tighter cluster than the controls, thus a 'metabolic lensing' effect is a result of the stressor (Cd^{2+}) having the largest overall effect on metabolism within the culture system.

The primary aim of this work was to explore the biochemical differences between control and dosed sample groups of *S. cucubalus* cell cultures following exposure to Cd^{2+}. A PCA scores plot following re-analysis using the control and dosed sample groups only is shown in Fig. 3. It can be seen that the groups are readily discriminated in PC1. Having obtained a model that is capable of discriminating between the two sample classes of interest, the dataset was interrogated in order to determine those variables, (and in turn NMR regions, and ultimately biochemical entities) that were most important in class separation. PCA produces a series of new variables (PCs) based on linear combinations of the original variables. By analyzing the weighting given to each of the original variables, i.e. the degree of correlation between the variables and the direction of the new model, it is possible to determine their importance, known as the variable loadings. As seen in Fig. 3, the separation between the control and dosed groups was achieved in PC1. It was, therefore, possible to determine variable importance by analyzing the correlation of each variable with PC1, Fig. 4. A positive value in the loadings plot shown in Fig. 4 implies a positive correlation with the scores in PC1. Thus all variables with positive values in Fig. 4 are positively correlated with the control group, whilst the variables with negative values are correlated with the dosed group. When the

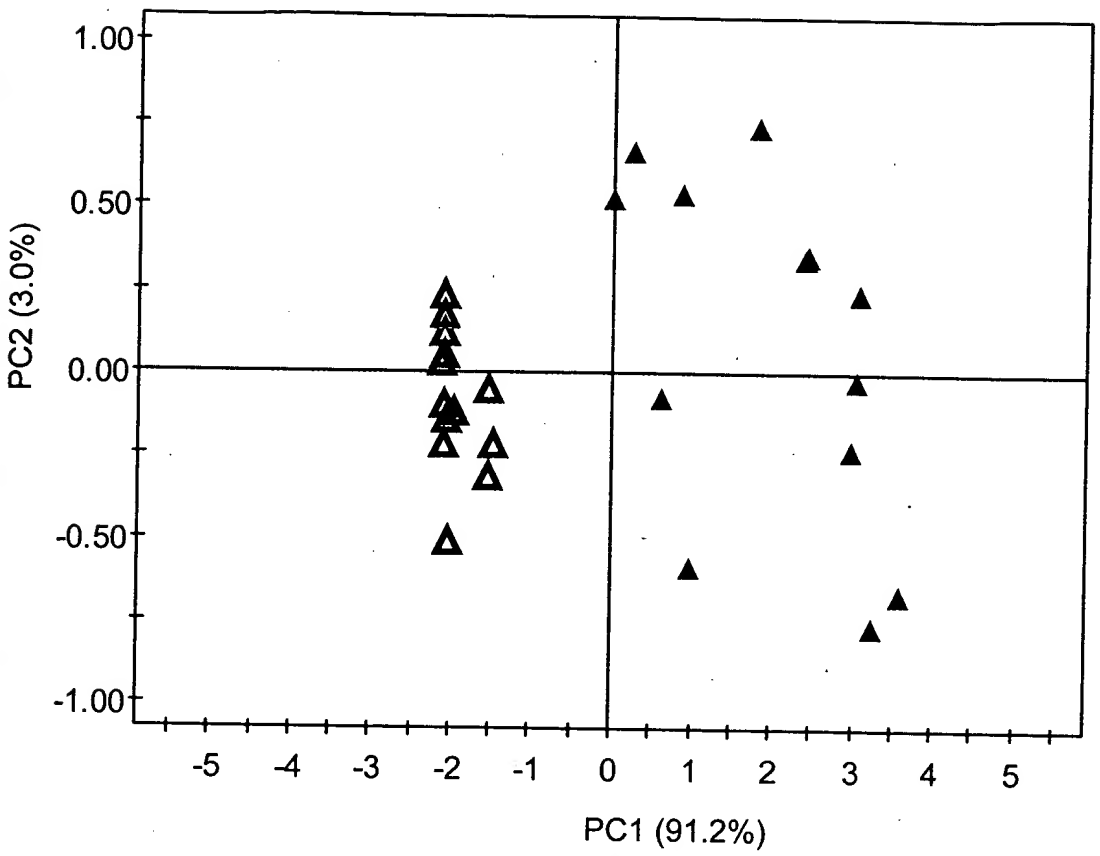


Fig. 3. Scores plot (PC1 v PC2) for dosed (open purple triangles) and control (closed blue triangles) samples groups following PC analysis. 94.2% of the variance within the dataset are explained in this plot.

variable loadings are plotted on the NMR frequency scale, it is apparent which NMR spectral regions are important. Hence by reference to established NMR assignments for small molecules and in certain cases 2-D NMR experiments (not shown), it is possible to identify the metabolite patterns that discriminated between the two groups.

The change that had the most influence on the discrimination between control and dosed groups was in the concentration of glutamine, which was substantially reduced between control and dosed groups, i.e. it has a large positive value in Fig. 4, indicating high levels in the control group, and lower levels in the dosed group. The major regions showing changes between control and dosed groups are summarized along with their metabolite assignments in Table 1. In general, the metabolites that were shown to be important are linked to the TCA cycle. Increased glucose levels suggests that utilization of glucose is reduced in Cd^{2+} exposed plants, while the presence of acetate may indicate either increased lipid metabolism or reduced utilisation of acetyl CoA in the TCA cycle. In addition, changes in levels of glutamate and malate may be related to changes in TCA intermediates. Although the anticipated presence of phytochelatins was not observed, this is due to the fact that they are present at too low a level for direct observation by ^{1}H NMR spectroscopy. This is particularly the case for a complex matrix like plant extracts where the dynamic range imposed by other metabolites places restrictions on otherwise observable species. The total amount of phytochelatins present in the dosed group, as determined by HPLC assay was approximately 1.5 μmol g^{-1} lyophilized material, with each phytochelatin present in the 50–1220 nmol g^{-1} lyophilized material range (data not shown; phytochelatins were not detected in either control or predose groups).

This work demonstrates that the combination of high resolution ^{1}H NMR spectroscopy with multivariate data analysis is readily amenable to the rapid screening of biological samples in order to produce a metabolic profile, which at its most basic level can allow metabolic fingerprints to be generated. Further, the implementation of chemometric approaches to interrogate the resulting complex data allows significant biochemical changes to be readily extracted from the data. By virtue of the NMR spectra already obtained, it is then possible to elucidate the nature of the metabolites that are key in the separation between sample groups.

While the more conventional analytical approach using GC–MS allows the detection and quantitation of many compounds during the execution of the chromatographic run, pre-analysis derivatization and thus pre-selection of the 'expected' metabolites prior to analysis poses an

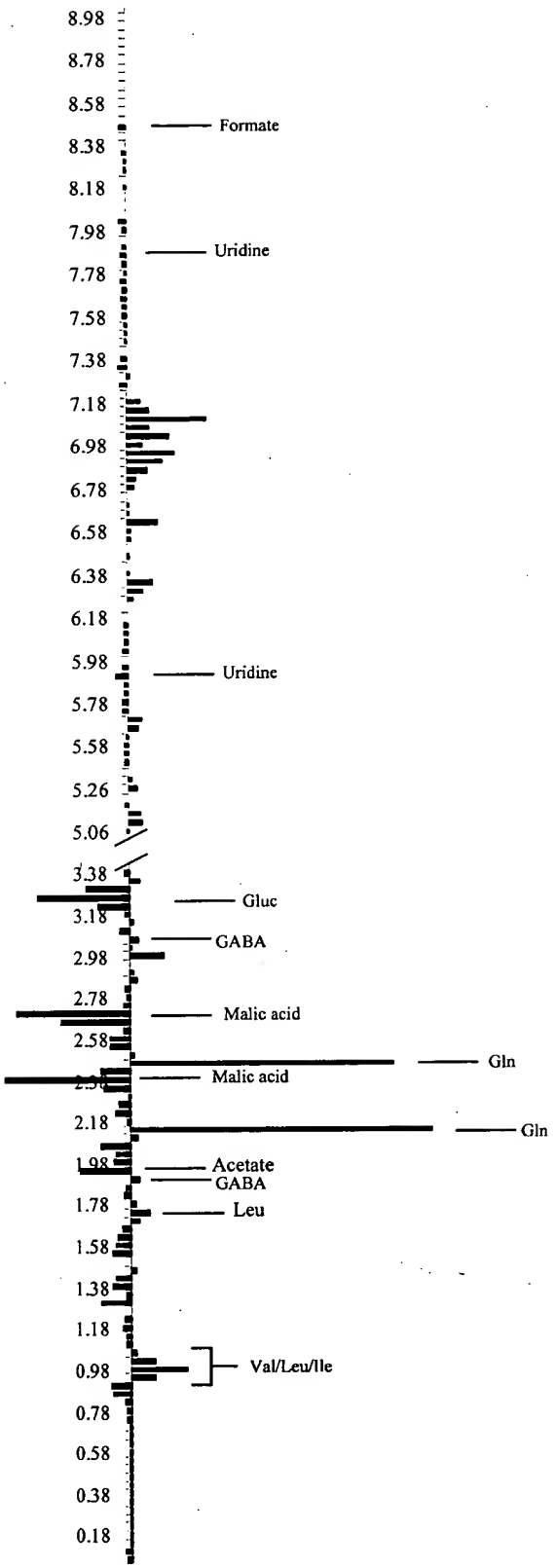


Fig. 4. Loadings column plot for dosed and control showing PC1 only. This plot allows elucidation of the chemical entities that are key in separating the control and dosed groups following PCA. Variables with a large positive value are positively correlated with the control group, whilst those with negative values are positively correlated with the dosed group.

obvious limitation of the methodology, as many non-derivatized chemical classes will be lost to the analysis. ^{1}H NMR spectroscopic approaches on the other hand, benefit from the non-selective nature of the technique, which means that no prior knowledge or judgement of the samples is required. NMR spectroscopy is an information rich technique providing key information for the structural identification of the metabolites detected. At the same time, this technique is not impeded by problems of differential detection between compound classes displaying differing chemical properties (such as ionisation in the case of mass spectrometric detection or UV absorbance in the case of HPLC for example). This means that NMR based approaches to metabolomics and metabonomics offer clear analytical advantages over alternative techniques, although NMR and MS approaches may also be considered in many applications to be complementary. In addition, the limited sample pre-treatment/derivatizations necessary, and the relatively short acquisition times mean that NMR spectroscopy may be utilized as a high throughput technique capable of rapidly analyzing the sample numbers required for statistically relevant studies.

With regards to this current work, it has been demonstrated that exposure of *S. cucubalus* cells to Cd^{2+} results in biochemical changes relating to energy production and the TCA cycle. There are indications however that lipid metabolism is also altered, perhaps in response to the down regulation of glucose metabolism. It may be hypothesized that it is this ability to switch the method of energy metabolism that imparts the Cd^{2+} tolerance to *S. cucubalus* whilst exposure to Cd^{2+} in highly sensitive species such as barley (*Hordeum vulgare*) results in reduced plant growth (Vassilev et al., 1998). In addition, while non-tolerant species show an increase in the levels of the stress biomarker proline (Vassilev et al., 1998), no increase in proline was observed in this study. Finally, this approach to metabolomic analysis has allowed the demonstration of the concept of 'metabolic lensing' with the variation within sample classes reduced between control and dosed classes as a result of the xenobiotic effect being greater than the inherent variance within a sample population. This suggests that it is important to obtain sufficient data points within a study to allow this phenomenon to be clearly identified as such, and also that biochemical variation is a factor that must be considered when planning metabolomic analyses.

3. Experimental

3.1. *S. cucubalus* suspension cell cultivation and sample preparation

Sterile *S. cucubalus* suspension cells (7 day old culture obtained from existing cultures at the Institute of Plan

Summary of the major changes between Cd^{2+} dosed and control sample groups

NMR spectral region (and intensity change between control and dosed)	Assignment	Concentration[a]/μmol/g dry weight (average, $n=3$)	
		Control	Dosed
Region 0.94–1.02 (decrease)	Valine	1.22±0.04	1.1±0.2
	Isoleucine	0.53±0.19	0.5±0.1
	Leucine	1.9±0.2	2.4±0.3
Region 1.94 (increase)	Acetate	14±2	17±4
Regions 2.14, 2.46 (decrease)	Glutamine	16.7±0.8	9.3±0.9
Region 2.38, 2.70, 2.66 (increase)	Malic acid	17±8	26±7
Region 3.26 3.22 4.34 (increase)	Glucose	[b]	[b]
Region 6.94–7.10 (decrease)	Unknown aromatic compounds	N/A	N/A

[a] Levels given are approximate only due to the overlap of resonances within the spectra, and the inherent errors associated with low level quantitation.
[b] Figures not given due to overlap of the glucose resonances.

Biochemistry, Halle, Germany) were vacuum filtered and washed with sterile water. A representative sample was flash frozen, lyophilized and taken as predose sample.

Two 1 l Erlenmeyer flasks with 250 ml fresh Linsmaier–Skoog growing media (Linsmaier and Skoog, 1965) were prepared, and 40 g (fr. wt.) cells added to each flask. In addition, one flask contained 3 ml sterile water (control flask), whilst the other flask contained 3 ml 12.5 mM $CdCl_2$ (final Cd^{2+} concentration 150 μM, dosed flask). Both flasks were cultivated under sterile conditions for 3 days (gyratory shaker 100 rpm, diffuse light 650 lux, 22 °C). After 3 days, cells from both flasks were vacuum filtered and washed with sterile water. Filtered cells were then flash frozen with liquid nitrogen and lyophilized.

Phytochelatin content of the cells was determined by HPLC with dithio-bis-nitrobenzoic acid postcolumn derivatization as described previously (Oven et al., 2002).

Replicates (approx 20 mg, $n=13$) of lyophilized cells from each flask were weighed out and added to D_2O (1 ml, containing 0.05% w/v 3-(trimethylsilyl) propionic-2,2,3,3-d_4 acid (sodium salt) (TSP) as NMR reference). Samples were agitated and then centrifuged at 13,000 rpm for 15 min. Supernatant (700 μl) was taken for NMR analysis.

3.2. ^1H NMR spectroscopy

NMR spectra were run on a Bruker (Bruker GmbH, Rheinstetten, Germany) DRX 600 Spectrometer, operating at 600.22 MHz for the ^1H frequency, fitted with a broadband inverse geometry probe. Spectra were the result of the summation of 64 free induction decays, with data collected into 32k datapoints, a spectral width of δ 14 and an acquisition time of 1.95 s. The water signal was suppressed using a standard 1D-presaturation pulse sequence (Nicholson et al., 1995). Prior to Fourier transformation, an exponential line broadening equivalent to 0.3 Hz was applied to the free induction decays and spectra were referenced to TSP at δ 0.00.

Quantitation was performed using a delay between pulses of 30 s to ensure full longitudinal relaxation. Concentrations were then calculated for each metabolite based on a known concentration of TSP.

3.3. Multivariate data analysis

One dimensional 600 MHz ^1H NMR spectra were reduced to 252 discrete chemical shift regions by digitisation to produce a series of sequentially integrated regions δ 0.04 in width between δ −0.02 and 9.98, using Bruker AMIX software (version 2.0, Bruker GmbH, Germany). The resulting data matrix was exported into Microsoft® Excel and selected regions removed, i.e. around the residual water signal (δ 4.54–4.98), sucrose (from the media solution, δ 5.46–5.38, 4.30–4.18, 4.10–3.42) and TSP (δ −0.02 to 0.02). The remaining 212 integral regions were normalized to the whole spectrum for subsequent Principal Components Analysis (PCA) (Eriksson et al., 1999).

PCA was performed using SIMCA-P 8.0 multivariate data analysis software (Umetrics, Sweden), with mean centring of the data preceding PCA. The output from the PCA analysis consisted of scores plots (giving an indication of the differentiation of the classes in terms of biochemical similarity), and loadings plots, which give an indication as to which NMR spectral regions were important with respect to the classification obtained in the scores plots.

Acknowledgements

The authors wish to thank the European Science Foundation Plant Adaptation Programme for providing financial assistance to NJCB to allow completion of this work. Work at Halle was also supported by SFB369 of Deutsche Forschungsgemeinschaft, Bonn, Germany.

References

Bailey, N.J., Sampson, J., Hylands, P.J., Nicholson, J.K., Holmes, E., 2002. Multi-component metabolic classification of commercial feverfew preparations via high field 1H NMR spectroscopy and chemometrics. Planta Med. 68, 1–5.

Belton, P.S., Colquhoun, I.J., Kemsley, E.K., Delgadillo, I., Roma, P., Dennis, M.J., Sharman, M., Holmes, E., Nicholson, J., Spraul, M., 1998. Application of chemometrics to the 1H NMR spectra of apple juices: discrimination between apple varieties. Food Chem. 61, 207–213.

Cobbett, C.S., 2000. Phytochelatin biosynthesis and function in heavy metal detoxification. Curr. Opin. Plant Biol. 3, 211–216.

di Toppi, L.S., Gabbrielli, R., 1999. Response to cadmium in higher plants. Env. Exp. Bot. 41, 105–130.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wold, S., 1999. Introduction to Multi- and Megavariate Data Analysis Using Projection Methods (PCA and PLS). Umetrics AB, Umeå, Sweden.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Tretheway, R.N., Willmitzer, L., 2000. Metabolite profiling for plant functional genomics. Nat. Biotech. 18, 1157–1161.

Gavaghan, C.L., Holmes, E., Lenz, E., Wilson, I.D., Nicholson, J., 2000. A NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences: application to the C57BL10J and Alpk:ApfCD mouse. FEBS Lett. 484, 169–174.

Griffin, J., Walker, L., Garrod, S., Holmes, E., Shore, R., Nicholson, J., 2000. NMR spectroscopy based metabonomic studies on the comparative biochemistry of the kidney and urine of the bank vole (*Clethrionomys glareolus*), wood mouse (*Apodemus sylvaticus*), white toothed shrew (*Crocidura suaveolens*) and the laboratory rat. Comp. Biochem. Phys. B 127, 357–367.

Grill, E., Winnacker, E.-L., Zenk, M., 1985. Phytochelatins: the principal heavy metal complexing peptides of higher plants. Science 230, 674–676.

Holmes, E., Nicholson, J., Tranter, G., 2001. Metabonomic characterization of genetic variations in toxicological and metabolic responses using probabilistic neural networks. Chem. Res. Toxicol. 14, 182–191.

Lindon, J.C., Holmes, E., Nicholson, J.K., 2001. Pattern recognition methods and applications in biomedical magnetic resonance. Prog. Nuc. Mag. Res. Spec. 39, 1–40.

Linsmaier, E.M., Skoog, F., 1965. Organic growth factor requirements of tobacco tissue cultures. Physiol. Plant 18, 100–127.

Lombi, E., Zhao, F., Dunham, S., McGrath, S., 2000. Cadmium accumulation in populations of *Thlaspi caerulescens* and *Thlaspi goesingense*. New Phytol. 145, 11–20.

Nicholson, J., Foxall, P.J., Spraul, M., Farrant, R.D., Lindon, J., 1995. 750 MHz 1H and ^{1}H–^{13}C NMR spectroscopy of human blood plasma. Anal. Chem. 67, 793–811.

Nicholson, J., Lindon, J., Holmes, E., 1999. Metabonomics: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 29, 1181–1189.

Nicholson, J.K., Connelly, J., Lindon, J.C., Holmes, E., 2002. Metabonomics: a platform for studying drug toxicity and gene function. Nat. Rev. Drug Disc. 1, 153–160.

Nicholson, J.K., Kendall, M.D., Osborn, D., 1983. Cadmium and mercury nephrotoxicity. Nature 304, 633–635.

Nicholson, J.K., Osborn, D., 1983. Kidney lesions in pelagic seabirds with high tissue-levels of cadmium and mercury. J. Zool. 200, 99–118.

Noteborn, H.P., Lommen, A., van der Jagt, R., Weseman, J.M., 2000. Chemical fingerprinting for the evaluation of unintended secondary metabolic changes in transgenic food crops. J. Biotech. 77, 103–114.

Oven, M., Page, J., Zenk, M., Kutchan, T.M., 2002. Molecular characterization of the homo-phytochelatin synthase of soybean Glycine max- relation to phytochelatin synthase. J. Biol. Chem. 277, 4747–4754.

Raamsdonk, L.M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M.C., Berden, J.A., Brindle, K.M., Kell, D.B., Rowland, J.J., Westerhoff, H.V., van Dam, K., Oliver, S.G., 2001. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nat. Biotech. 19, 45–50.

Roessner, U., Wagner, C., Kopka, J., Tretheway, R.N., Willmitzer, L., 2000. Simultaneous analysis of metabolites in potato tuber by GCMS. Plant J. 23, 131–142.

Roessner, U., Willmitzer, L., Fernie, A.R., 2001. High resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. Plant Physiol. 127, 749–764.

Vassilev, A., Berova, M., Zlatev, Z., 1998. Influence of Cd on growth, chlorophyll content, and water relations in young barley plants. Biol. Plant 41, 601–606.

Zenk, M., 1996. Heavy metal detoxification in higher plants—review. Gene 179, 21–30.

Mellerson, Kendra

From: Gakh, Yelena
Sent: Tuesday, August 05, 2003 2:33 PM
To: STIC-EIC1700
Subject: 09890973

Dear Kendra:

please order one more list:

9. TITLE: "Metabonomics": understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data"
AUTHOR(S): *Nicholson, J. K.; Lindon, J. C.; Holmes, E.*
CORPORATE SOURCE: Biological Chemistry, Biomedical Sciences Division, Imperial College of Science, Technology and Medicine, University of London, London, SW7 2AZ, UK
SOURCE: **Xenobiotica (1999), 29(11), 1181-1189**

Thank you,

Yelena

Biotech
QP501.A31
QR Adams

Yelena G. Gakh, Ph.D.

Patent Examiner
USPTO, cp3/7B-08
(703)306-5906

Article title	'Metabonomics': Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data
Article identifier	0049825499105956
Authors	Nicholson_J_K Lindon_J_C Holmes_E
Journal title	Xenobiotica
ISSN	0049-8254
Publisher	Taylor and Francis UK
Year of publication	1999
Volume	29
Issue	11
Supplement	0
Page range	1181-1189
Number of pages	9
User name	Adonis
Cost centre	
PCC	$19.00
Date and time	Wednesday, August 06, 2003 4:18:18 AM

18069302
XENOBIOTICA, 1999, VOL. 29, NO. 11, 1181–1189

'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data

J. K. NICHOLSON*, J. C. LINDON and E. HOLMES

Biological Chemistry, Biomedical Sciences Division, Imperial College of Science, Technology and Medicine, University of London, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ, UK

Received 5 July 1999

Introduction

The rapid evolution of drug discovery science, fuelled by combinatorial library-based synthesis programmes, has led to increased pressure on the drug safety evaluation process. Once potential drugs have passed the primary biological screening procedures, losses of drug candidate compounds from the product development pipeline (known as 'attrition') need to be minimized. Hence, there is an intensive search for new analytical technologies that will maximize efficiency of lead compound selection based both on efficacy and safety and will minimize overall attrition rates. Current bioanalytical approaches include measurements of responses of living systems to drugs either at the genetic level or at the level of expression of cellular proteins, using so-called genomic and proteomic methods respectively. At present both genomics and proteomics are expensive and labour-intensive, yet potentially are powerful tools for studying different levels of the biological response to xenobiotic exposure. However, even in combination, genomics and proteomics do not provide the range of information needed for an understanding of the integrated cellular function in living systems, since both ignore the dynamic metabolic status of the whole organism. Thus, a new NMR-based 'metabonomic' approach is proposed that is aimed at the augmentation and complementation of the information provided by measuring the genetic and proteomic responses to xenobiotic exposure. Metabonomics is defined as 'the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification'. This concept has arisen from work on the application of ^{1}H-NMR spectroscopy to study the multicomponent metabolic composition of biofluids, cells and tissues over the past two decades (e.g. Nicholson *et al.* 1983, 1985, Bales *et al.* 1984, Gartland *et al.* 1989, Nicholson and Wilson 1989, Moka *et al.* 1998). Also studies utilizing pattern recognition (PR), expert systems and related bio-informatic tools are used to interpret and classify complex NMR-generated metabolic data sets (Gartland *et al.* 1991, Holmes *et al.* 1992, 1994, 1998a, b, Anthony *et al.* 1994, Spraul *et al.* 1997, Beckwith-Hall *et al.* 1998). There is also a significant background to this work in other research fields, notably metabolic control analysis (Kacser and Burns 1973, Kacser 1993, Goodacre *et al.* 1996), and there is a related concept of the 'Metabolome' that represents the total small molecule complement of a cell. However, metabonomics deals with detecting,

* Author for correspondence.

Xenobiotica ISSN 0049-8254 print/ISSN 1366-5928 online © 1999 Taylor & Francis Ltd
http://www.tandf.co.uk/JNLS/xen.htm http://www.taylorandfrancis.com/JNLS/xen.htm

identifying, quantitating and cataloguing the *history* of time-related metabolic changes in an integrated biological system rather than the individual cell. Such multidimensional metabolic trajectories are then related to the biological events in an ongoing pathophysiological process. Here, provided is a brief background to the useful properties of metabonomic data sets and the possible uses of NMR-based metabonomics for toxicological classification and biomarker or surrogate marker identification *in vivo*.

Genomic and proteomic approaches to drug toxicity assessment

Development of new tools in structural molecular biology has led to an increased understanding of the organization of the genome. This knowledge combined with a massive increase in the ability to identify and sequence genes has led to the point where the entire genome of > 20 prokaryotic organisms, e.g. *Archaeoglobus fulgidus* (Klenk *et al.* 1997), has already been sequenced together with one eukaryotic organism with ~ 19000 genes and $> 93 \times 10^6$ bp (*Caenorhabditis elegans*; The *C. elegans* Sequencing Consortium 1998). A complete description of the human genome with ~ 80000 genes is probably only a few years away. One of the intellectual products of the molecular biology revolution has been the concept of 'genomics', which is basically a semiquantitative approach to the measurement of gene expression. In the context of drug discovery and for the purposes of toxicological assessment, the genomic approach involves the observation of altered gene expression after drug exposure. The technology involves a new generation of proprietary 'gene chips', which are small disposable devices encoded with an array of genes that respond to extracted cellular mRNA produced after exposure to a foreign compound which has caused the 'switching on' of various genes (Sinclair 1999). Many genes can be placed on a chip array and patterns of gene switching caused by xenobiotic exposure can be monitored rapidly in this way, although at some considerable cost. However, relationships between gene regulation/expression and the integrated function and control of cellular systems (so-called *functional* genomics) are still far from clear, and will remain so for many years after the complete sequencing of the human genome. The main reason for this is that the vast majority of DNA is non-coding, yet protein coding sequences or genes cannot function as isolated units and can require the presence of neighbouring genes and/or non-coding DNA. The lack of understanding of the biological consequences of altered gene expression has led to the development of proteomics, which is concerned with the semiquantitative measurement of the production of cellular proteins in response to drug exposure and other pathophysiological processes (Anderson *et al.* 1996, Aicher *et al.* 1998, Geisow 1998). Proteomic measurements utilize a variety of technologies, but all involve a protein separation method, e.g. 2D gel-electrophoresis, allied to a chemical characterization method, usually, some form of mass spectrometry (MS). While potentially less expensive than genomics, proteomics is very slow and labour-intensive at present. More importantly, although these measurements may ultimately give profound insights into toxicological mechanisms and provide new surrogate biomarkers of disease, at present it is very difficult to relate genomic and proteomic findings to classical indices of toxicity or toxicological end-points. One simple reason for this is that the current technology and approach precludes the measurement of a detailed time-course of the response to drug exposure or the measurement of responses in a multi-organ system. This may be particularly important for the many known

cases where the metabolism of the compound is a prerequisite for toxicity and especially true where the target organ is not the site of primary metabolism. An example is the case of compounds that form glutathione S-conjugates in the liver that are subsequently processed by β-lyase thus generating reactive intermediates that show ultimate target organ toxicity in the renal proximal tubules (Elfarra *et al.* 1986). There is a need for the development of novel methods that give information of *in vivo* multi-organ functional integrity in real time. NMR-based metabonomics offers one such approach to the generation of this type of information.

NMR-based metabonomics

Foreign compounds may interact with tissue and extracellular components of an animal at a series of organizational levels ranging from changes in genetic expression through protein production and integrated cellular biochemical regulation and control. In such cases there will be alterations detectable at all levels of biomolecular organization and a complete approach to the description of these changes might be termed as 'bionomics' (proposed by Professor Ian D. Wilson). In many cases, drugs exert their toxic effects by interacting directly with genetic material or by inducing the synthesis of drug metabolizing enzymes, which generate toxic products. In such cases genomic and proteomic approaches to toxicity assessment may be useful. However, xenobiotics may act only at the pharmacological level and, hence, may not affect gene regulation or expression. Also significant toxicological effects may be completely unrelated to gene switching or protein synthesis. Exposure to ethanol *in vivo* may switch on many genes, but this does not explain drunkenness! Hence, in many cases facile consideration of genomic and proteomic responses are likely to be ineffective at predicting drug toxicity. However, all drug-induced pathophysiological perturbations result in disturbances in the ratios and concentrations, binding or fluxes of endogenous biochemicals, either by direct chemical reaction or by binding to key enzymes or nucleic acids that control metabolism. If these disturbances are of sufficient magnitude, toxic effects will result that will affect the efficient functioning of the whole organism. In body fluids, metabolites are in dynamic equilibrium with those inside cells and tissues and, consequently, abnormal cellular processes in tissues of the whole organism following a toxic or metabolic insult will be reflected in altered biofluid compositions. In all cases the analytical problem usually involves the detection of 'trace' amounts of analytes in a very complex matrix with many potential interferences. It is critical, therefore, to choose a suitable analytical technique for the particular class of analyte of interest in the biomatrix, for example blood, plasma, urine, bile or organ samples. High-resolution ^{1}H-NMR spectroscopy appears particularly appropriate for investigating abnormal body fluid compositions as a wide range of metabolites can be quantified simultaneously with no sample preparation and 'without prejudice'. Other techniques such as MS may also be useful for generating metabolic data, but differential ionization efficiency in the complex could affect detectability and quantitiation. NMR spectroscopy may also be used effectively to screen for abnormal metabolite profiles in tissue extracts or cell suspensions. It has also been shown that the same approach can be used to investigate the metabolic composition of *intact* tissues using high-resolution magic angle spinning ^{1}H-NMR spectroscopy (Moka *et al.* 1998).

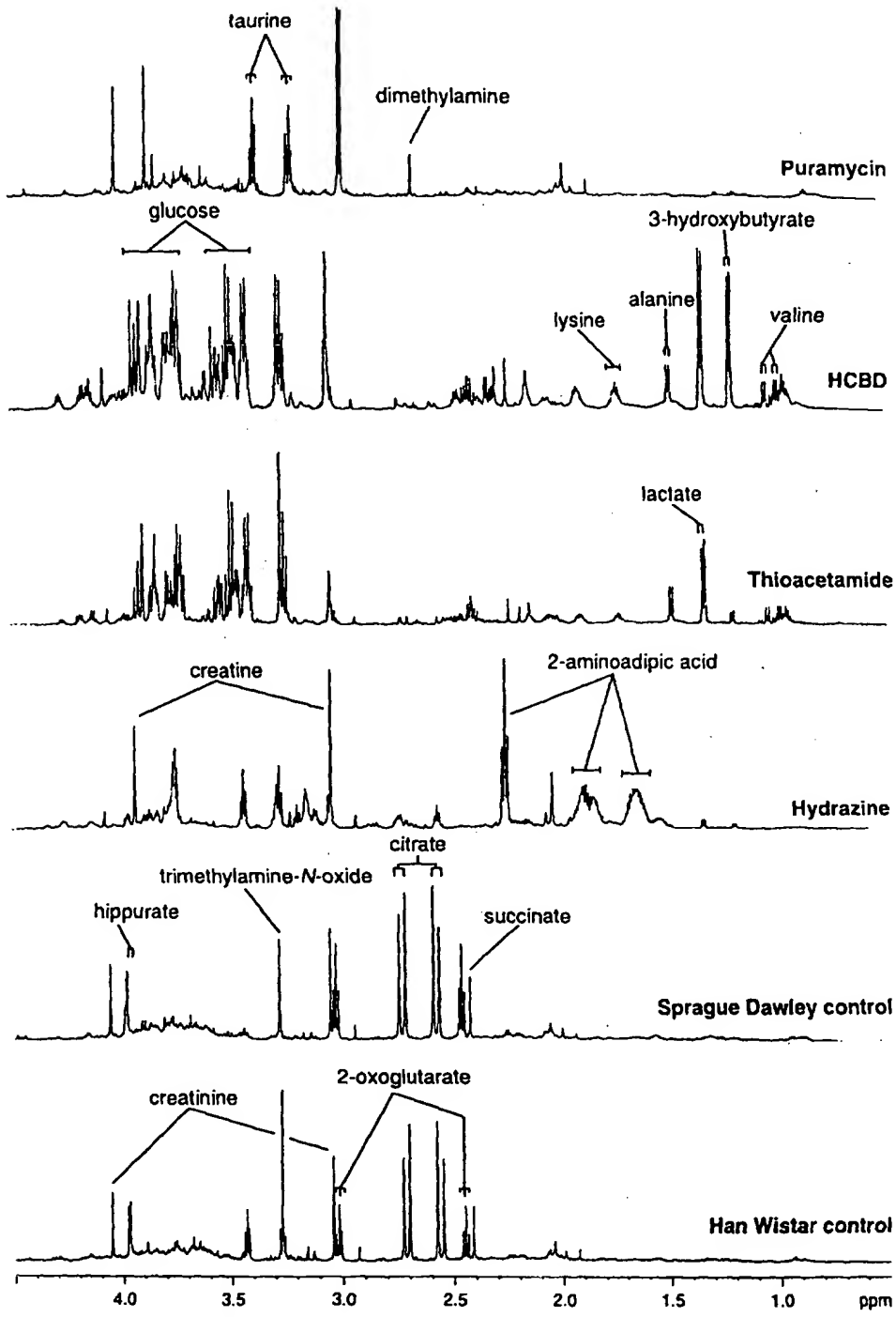


Figure 1. Partial 600 MHz ^{1}H-NMR spectra of a series of urines from the control rat, and those collected 8–24 h after treatment with various model toxins. HCBD, hexachloro-1,3-butadiene.

The exact pattern of endogenous metabolites in body fluids as detected by ^{1}H-NMR spectroscopy depends strongly on the type of toxin to which an animal has been exposed (Nicholson *et al.* 1983, 1985, Bales *et al.* 1984, Gartland *et al.* 1989, Nicholson and Wilson 1989). Each class of toxin produces characteristic changes in the concentrations and patterns of endogenous metabolites in biofluids and this provides information on the sites and basic mechanisms of the toxic process. A typical series of spectra from urine of rat treated with different toxins are shown in figure 1. Bio-analytically, the processes of generating such information is highly efficient, taking only a few minutes per sample and requiring little or no sample pretreatment or reagents. The spectra are very similar in the case of controls (two common models the Han Wistar and Sprague Dawley being shown), but different toxins cause characteristic metabolic perturbations. Because nearly all major classes of metabolic intermediate have characteristic NMR spectra, the technique is very useful for fingerprinting toxin-induced metabolic variations. Thus, ^{1}H-NMR spectroscopic analysis of biofluids has successfully uncovered numerous novel metabolic biomarkers of organ-specific toxicity in the rat, and it is in this 'exploratory' role that NMR as an analytical biochemistry technique excels. For example, changes in the levels of trimethylamine-*N*-oxide, *N*,*N*-dimethylglycine, dimethylamine and succinate are indicative of damage to the renal papilla for which no biochemical biomarkers existed previously (Gartland *et al.* 1989, 1991). Other urinary markers uncovered by ^{1}H-NMR urinalysis include taurine and creatine, which have been correlated with acute liver and testicular toxicity respectively (Nicholson *et al.* 1989, Gray *et al.* 1990, Sanins *et al.* 1990). Similar approaches can be used using 2D NMR spectroscopy (Nicholson and Wilson 1989). However, the biomarker information in NMR spectra of biofluids is much more subtle and rich than this, as hundreds of compounds representing many pathways can often be measured simultaneously, and it is the overall metabonomic response to toxic insult (occurring over time) that so well characterizes the lesion (Beckwith-Hall *et al.* 1998, Holmes *et al.* 1998a). The most efficient way to investigate these complex multiparametric data is to continue the 1D and 2D NMR metabonomic approach with PR methods.

Pattern recognition and expert system analysis of NMR-generated metabonomic data

A limiting factor in understanding the biochemical information from both 1D and 2D NMR spectra of tissues and biofluids is their very complexity; even 1D ^{1}H-NMR spectra (at 600 MHz or above) of biofluids may contain several thousand resolved lines. The NMR spectrum of a sample under study can be considered as an *n*-dimensional object the dimensions of which could be the concentrations of individual measurable metabolites or more simply the spectral intensity distribution. Thus, the NMR spectrum of the biofluid or tissue provides an *n*-dimensional metabolic fingerprint of the organism based on the sample studied, and this metabolic profile is characteristically changed according to the disease or toxic process. Hence, computer-based PR and expert system approaches have been used to interpret the NMR data obtained in various experimental toxicity states (Gartland *et al.* 1991, Holmes *et al.* 1992, 1994, 1998a, b, Anthony *et al.* 1994, Spraul *et al.* 1997, Beckwith-Hall *et al.* 1998). These statistical tools are very similar to those currently being explored by those in the fields of genomics and proteomics. The

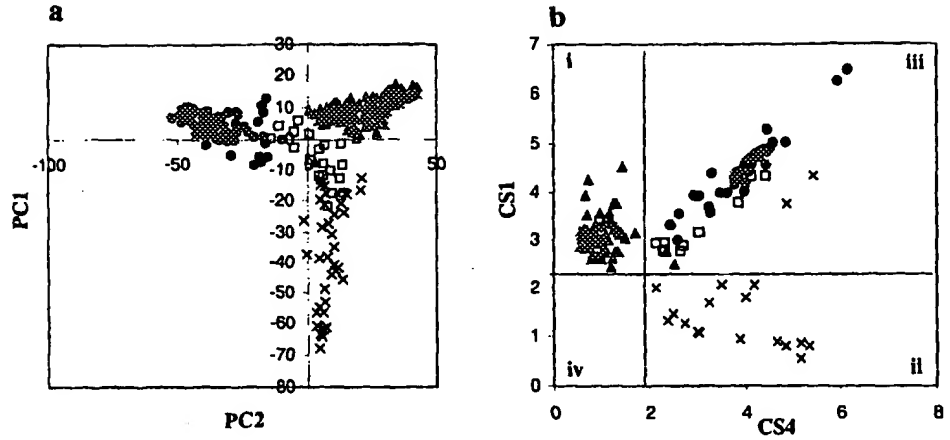


Figure 2. (a) Principal components map of data obtained from rat urines after treatment with lead acetate (□) hydrazine (×) and renal proximal tubular toxins affecting the S3 region (●) and controls (▲). (b) Cooman's residuals plot of test data set using a SIMCA model previously 'trained' using the same spectra shown in (a). Quadrant (i) shows samples unambiguously classified as controls, quadrant (ii) shows 'pure' hydrazine-toxicity classification, quadrant (iii) shows spectra from animals classified as neither control nor hydrazine-treated type, and quadrant (iv) shows an unoccupied field that would indicate mixed hydrazine-toxicity and control classification. In this example, two hydrazine-treated data points are misclassified and two controls are also misclassified as abnormal samples. The lines show the 95% confidence limits of the classifications based on the training set data.

simplest approach is to treat the NMR signal intensity data as a multi-sample array of metabolite concentration or excretion rate scores; it is not necessary to assign the spectrum at this stage as it is treated solely as a statistical object. PR is a general term applied to methods of data analysis that can be used to generate scientific hypotheses as well as *testing* hypotheses by reducing mathematically the many parameters. One of the most useful and easily applied PR techniques is principal components analysis (PCA). Principal components (PC) are new variables created from linear combinations of the starting variables with appropriate weighting coefficients. The properties of these PC are such that (1) each PC is orthogonal (uncorrelated) with all other PC and (2) the first PC contains the largest part of the variance of the data set (information content) with subsequent PC containing correspondingly smaller amounts of variance. Thus, a plot of the first two or three PC gives the 'best' representation, in terms of biochemical variation in the data set in two or three dimensions. Such PC maps can be used to visualize inherent clustering behaviour for drugs and toxins acting on each organ according to toxic mechanism (Nicholson and Wilson 1989, Gartland *et al.* 1991). Such an application of PCA to toxicological mapping of NMR-generated metabonomic data is shown in figure 2a in which there is distinct clustering of data points from the urines of individual animals exposed to different toxins. The position on a PC plot of a sample from a xenobiotic-treated animal is determined purely by its metabolic response as opposed to any other independent knowledge of the compound action; hence, the method is termed 'unsupervised'. Of course, the clustering information might be in lower PC and this also has to be examined. In this simple metabonomic approach a sample from an animal treated with a compound of unknown toxicity is compared with a database of NMR-generated metabolic data and its topographical fit on the PR map is determined (Holmes *et al.* 1998a, b). However, in the real world, toxicological data

are more complex as lesions develop and resolve in real time and, hence, there are time-related changes in NMR-detected metabolic profile (Holmes *et al.* 1992, Beckwith-Hall *et al.* 1998). Also, it is more rigorous to compare effects of xenobiotics in the original *n*-dimensional NMR metabonomic space. Hence, as an alternative approach and to develop automatic toxicity classification methods, it has proved efficient to use a 'supervised' approach to NMR data analysis. Here, a 'training set' of NMR metabonomic data is used to construct a mathematical model that predicts correctly the class of each sample. This training set is then tested with independent data ('test set') to determine the robustness of the computer-based model. These models are sometimes termed expert systems, but may comprise systems based on a range of different mathematical procedures such as principal components, artificial neural networks and rule induction. In all cases the methods allow the quantitative description of the multivariate boundaries that characterize and separate each class of xenobiotic in terms of their metabolic effects. Certain supervised methods, such as SIMCA (soft independent modelling of class analogy; Kowalski *et al.* 1986) also allow a level of probability to be placed on the goodness of fit. Using such systems a sample can be classified as belonging to a single class of toxicity, to multiple classes of toxicity (more than one target organ) or to no class. The latter case would indicate deviation from normality (control) based on the training set model but having a dissimilar metabolic effect to any toxicity class modelled in the training set (unknown toxicity type). An example of an expert systems based classification of toxicity data is shown in figure 2b. In this simple illustrative case SIMCA models were constructed for both control rat urines and for rat urines from hydrazine-dosed animals using a training set of NMR data. The Cooman's residuals plot shown in figure 2b demonstrates that the majority of the test controls and test hydrazine-treated spectra are correctly classified and S3 type renal cortical toxins and lead acetate (which causes a range of renal, haemopoeitic and hepatotoxic effects) are all correctly classified as neither control nor hydrazine type. By building an exhaustive series of models it is possible to use SIMCA and other methods to provide classification probabilities for a wide range of toxicity types.

The metabonomic expert systems currently under construction in our group can be considered to operate at three distinct levels of pathophysiological discrimination:

1. Classification of the sample or organism as 'normal or abnormal' according to metabonomic criteria derived from a large database of controls (this will be a useful tool in the control of NMR spectrometer automation using sequential flow injection NMR spectroscopy; Spraul *et al.* 1997).
2. Classification of the target organ for toxicity and site of action within the tissue.
3. Identification of the biomarkers of toxic effect and toxic mechanism classification for the compound under study.

Interestingly, these levels of classification or discrimination would also apply even if data were derived from genomic or proteomic studies and similar arguments could be applied to clinical diagnostic screening procedures. As the size of toxicological databases increases together with improvements in rapid throughput of NMR samples (300 samples per day per spectrometer is now possible with the first generation flow injection systems), more subtle expert systems will be necessary using techniques such as 'fuzzy logic', which permits greater flexibility in decision boundaries between classes. Using the metabonomic methods described above, it has already been possible to develop a prototype expert system for classification

at level 1, and has also effected level 2 classification procedures for a range of toxicological endpoints and target organs. The level 3 classification poses more complex problems in terms of expert system development, but detailed biomarker information can already be obtained from inspection of the PC loadings (Holmes *et al.* 1998b).

In conclusion, there is a vast range of biochemical, toxicological and clinical chemical problems that can be addressed using metabonomics based on high-resolution ^{1}H-NMR spectroscopy of biomaterials. At present even simple ^{1}H-NMR experiments on whole biofluids can generate substantial amounts of metabolic data that can give surprisingly detailed insight into the biochemical processes in the whole organisms and the investigation of species differences in terms of toxicological biomarkers. The numbers of applications of metabonomics is bound to increase in parallel with ongoing developments in instrumentation and techniques. In particular, the development of computer-based PR and expert systems for data analysis is expected to make major contributions to the advancement of NMR-based metabolic science. Other important areas accessible to metabonomic investigation include studies on biochemical consequences of genetic modification, e.g. in 'knock-out animals', investigations into effects of environmental pollutants, for clinical evaluation of drug therapy and efficacy, and the investigation of idiosyncratic toxicity in man. Finally, it should soon be possible to combine genomic, proteomic and metabonomic data sets into comprehensive 'bionomic' systems for the holistic evaluation of perturbed *in vivo* function.

Acknowledgements

The authors thank Professor Ian Wilson, AstraZeneca Pharmaceuticals, Alderley Park, (who also coined the term 'bionomics') and Professor Jeremy Everett, Pfizer Pharmaceuticals, Sandwich, for helpful discussions on the background and basic philosophy of the work on metabonomics over several years.

References

Aicher, L., Wahl, D., Arce, A., Grenet, O. and Steiner, S., 1998, New insights into cyclosporine A nephrotoxicity by proteome analysis. *Electrophoresis*, **19**, 1998–2003.

Anderson, N. L., Taylor, J., Hofmann, J. P., Esquer-Blasco, R., Swift, S. and Anderson, N. G., 1996, Simultaneous measurement of hundreds of liver proteins: application in assessment of liver function. *Toxicologic Pathology*, **24**, 72–76.

Anthony, M. L., Sweatman, B. C. Beddell, C. R., Lindon, J. C. and Nicholson, J. K., 1994, Pattern recognition classification of the site of nephrotoxicity based on metabolic data derived from high resolution proton nuclear magnetic resonance spectra of urine. *Molecular Pharmacology*, **46**, 199–211.

Bales, J. R., Higham, D. P., Howe, I., Nicholson, J. K. and Sadler, P. J., 1984, Use of high resolution proton nuclear magnetic resonance spectroscopy for rapid multi-component analysis of urine. *Clinical Chemistry*, **30**, 426–432.

Beckwith-Hall, B. M., Nicholson, J. K., Nicholls, A., Foxall, P. J. D., Lindon, J. C., Connor, S. C., Abdi, M., Connelly, J. and Holmes, E., 1998, Nuclear magnetic resonance spectroscopic and principal components analysis investigations into biochemical effects of three model hepatotoxins. *Chemical Research in Toxicology*, **11**, 260–272.

Elfarra, A. A., Jakobson, I. and Anders, M. W., 1986, Mechanism of S-(1,2-dichlorovinyl) glutathione-induced nephrotoxicity *Biochemical Pharmacology*, **35**, 283–288.

Gartland, K. P. R., Beddell, C., Lindon, J. C. and Nicholson, J. K., 1991, The application of pattern recognition methods to the analysis and classification of toxicological data derived from NMR spectroscopy of urine. *Molecular Pharmacology*, **39**, 629–642.

Gartland, K. P. R., Bonner, F. and Nicholson, J. K., 1989, Investigations into the biochemical effects of region-specific nephrotoxins. *Molecular Pharmacology*, **35**, 242–251.

Geisow, M. J., 1998, Proteomics: one small step for a digital computer, one giant leap for humankind. *Nature Biotechnology*, **16**, 206.

GOODACRE, R. RISCHERT, D. J., EVANS, P. M. and KELL, D. B., 1996, Rapid authentication of animal cell lines using pyrolysis mass spectrometry and autoassociative artificial neural networks. Cytotechnology, **21**, 231–241.

GRAY, J., NICHOLSON, J. K., CREASY, D. M. and TIMBRELL, J. A., 1990, Studies on the relationship between testicular toxicity and urinary and plasma creatine concentration. *Archives of Toxicology*, **64**, 443–450.

HOLMES, E., BONNER, F. W., SWEATMAN, B. C., LINDON, J. C., BEDDELL, C. R., RAHR, E. and NICHOLSON, J. K., 1992, NMR spectroscopy and pattern recognition analysis of the biochemical processes associated with the progression and recovery from nephrotoxic lesions in the rat induced by mercury (II) chloride and 2-bromoethanamine. *Molecular Pharmacology*, **42**, 922–930.

HOLMES, E., FOXALL, P. J. D., NICHOLSON, J. K., NEILD, G. H., BROWN, S. M. BEDDELL, C., SWEATMAN, B. C., RAHR, E., LINDON, J. C., SPRAUL, M. and NEIDIG, P., 1994, Automatic data reduction and pattern recognition methods for analysis of ^{1}H nuclear magnetic resonance spectra of human urine form normal and pathological states. *Analytical Biochemistry*, **220**, 284–296.

HOLMES, E., NICHOLLS, A. W., LINDON, J. C., RAMOS, S., SPRAUL, M., NEIDIG, P., CONNOR, S. C., CONNELLY, J., DAMMENT, S. J. P., HASELDEN, J. N. and NICHOLSON, J. K., 1998a, Development of a model for classification of toxin-induced lesions using ^{1}H-NMR spectroscopy of urine combined with pattern recognition. *NMR in Biomedicine*, **11**, 1–10.

HOLMES, E., NICHOLSON, J. K., NICHOLLS, A. W., LINDON, J. C., CONNOR, S. C., POLLY, S. and CONNELLY, J., 1998b, Identification of novel biomarkers of renal toxicity using automatic data reduction techniques and PCA of proton NMR spectra of urine. *Chemometrics and Intelligent Laboratory Systems*, **44**, 251–261.

KACSER, H., 1993, Recent developments beyond metabolic control analysis. *Biochemical Society Transactions*, **23**, 387–391.

KACSER, H. and BURNS, J. A., 1973, The control of flux. In D. D. Davies (ed.), *Rate Control of Biological Processes. Symposium of the Society for Experimental Biology*, Vol. 27 (Cambridge: Cambridge University Press), pp. 65–104.

KLENK, H. P. *et al.*, 1997, The complete genome sequence of the hyperthermophilic, sulfate-reducing archaeon *Archaeoglobus fulgidus*. *Nature*, **390**, 364–370.

KOWALSKI, B., SHARAF, D. and ILLMAN, D., 1986, *Chemometrics* (New York: Wiley).

MOKA, D., VORREUTHER, R., SHICHA, H., HUMPFER, E., LIPINSKI, M., SPRAUL, M., FOXALL, P. J. D., NICHOLSON, J. K. and LINDON, J. C., 1998, Biochemical classification of kidney carcinoma biopsy samples using magic angle spinning ^{1}H-NMR spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, **17**, 125–132.

NICHOLSON, J. K., BUCKINGHAM, M. J. and SADLER, P. J., 1983, High resolution proton NMR studies of vertebrate blood and plasma. *Biochemical Journal*, **211**, 605–615.

NICHOLSON, J. K., HIGHAM, D., TIMBRELL, J. A. and SADLER, P. J., 1989, Quantitative ^{1}H-NMR urinalysis studies on the biochemical effects of acute cadmium exposure in the rat. *Molecular Pharmacology*, **36**, 398–404.

NICHOLSON, J. K., TIMBRELL, J. A. and SADLER, P. J., 1985, Proton NMR spectra of urine as indicators of renal damage: Mercury nephrotoxicity in rats. *Molecular Pharmacology*, **27**, 644–651.

NICHOLSON, J. K. and WILSON, I. D., 1989, High resolution proton NMR spectroscopy of biological fluids. *Progress in NMR Spectroscopy*, **21**, 449–501.

SANINS, S. M., TIMBRELL, J. A., ELCOMBE, C. R. and NICHOLSON, J. K., 1990, Hepatotoxin-induced hypertaurinuria: a proton NMR study. *Archives of Toxicology*, **64**, 407–411.

SINCLAIR, B., 1999, Everything 's great when it sits on a chip: a bright future for DNA arrays. *The Scientist*, **13**, 18–20.

SPRAUL, M., HOFMANN, M., ACKERMANN, M., NICHOLLS, A. W., DAMMENT, S. J. P., HASKELDEN, J. N., SHOCKCOR, J. P., NICHOLSON, J. K. and LINDON, J. C., 1997, Flow injection ^{1}H-NMR spectroscopy combined with pattern recognition: implications for rapid structural studies and high throughput biochemical screening. *Analytical Communications*, **34**, 339–341.

THE *C. elegans* SEQUENCING CONSORTIUM, 1998, Genome sequence of the nematode *C. elegans*. *Science* [special issue], **11 December**, 2041-2046.

Mellerson, Kendra

From: Gakh, Yelena
Sent: Tuesday, August 05, 2003 2:33 PM
To: STIC-EIC1700
Subject: 09890973

Dear Kendra:

please order one more list:

3. TITLE: "Metabonomics classifies pathways affected by bioactive compounds. Artificial neural network classification of NMR spectra of plant extracts"
AUTHOR(S): *Ott, Karl-Heinz; Aranibar, Nelly; Singh, Bijay; Stockton, Gerald W.*
CORPORATE SOURCE: BASF Agro Research, Princeton, NJ, 08543, USA
SOURCE: **Phytochemistry (Elsevier) (2003), 62(6), 971-985**

[redacted]

Thank you,

Yelena

Biotech
QC861.P45

Yelena G. Gakh, Ph.D.

Patent Examiner
USPTO, cp3/7B-08
(703)306-5906

Available online at www.sciencedirect.com
SCIENCE DIRECT®
PHYTOCHEMISTRY
Phytochemistry 62 (2003) 971–985
www.elsevier.com/locate/phytochem

Metabonomics classifies pathways affected by bioactive compounds. Artificial neural network classification of NMR spectra of plant extracts

Karl-Heinz Ott*,[1], Nelly Aranibar*,[2], Bijay Singh[3], Gerald W. Stockton[4]

BASF Agro Research, Princeton, NJ 08543, USA

Received 11 November 2002; received in revised form 19 November 2002

Abstract

The biochemical mode-of-action (MOA) for herbicides and other bioactive compounds can be rapidly and simultaneously classified by automated pattern recognition of the metabonome that is embodied in the ^{1}H NMR spectrum of a crude plant extract. The ca. 300 herbicides that are used in agriculture today affect less than 30 different biochemical pathways. In this report, 19 of the most interesting MOAs were automatically classified. Corn (*Zea mays*) plants were treated with various herbicides such as imazethapyr, glyphosate, sethoxydim, and diuron, which represent various biochemical modes-of-action such as inhibition of specific enzymes (acetohydroxy acid synthase [AHAS], protoporphyrin IX oxidase [PROTOX], 5-enolpyruvylshikimate-3-phosphate synthase [EPSPS], acetyl CoA carboxylase [ACC-ase], etc.), or protein complexes (photosystems I and II), or major biological process such as oxidative phosphorylation, auxin transport, microtubule growth, and mitosis. Crude isolates from the treated plants were subjected to ^{1}H NMR spectroscopy, and the spectra were classified by artificial neural network analysis to discriminate the herbicide modes-of-action. We demonstrate the use and refinement of the method, and present cross-validated assignments for the metabolite NMR profiles of over 400 plant isolates. The MOA screen also recognizes when a new mode-of-action is present, which is considered extremely important for the herbicide discovery process, and can be used to study deviations in the metabolism of compounds from a chemical synthesis program. The combination of NMR metabolite profiling and neural network classification is expected to be similarly relevant to other metabonomic profiling applications, such as in drug discovery.

Keywords: Acetochlor; Amitrole; Artificial intelligence; Benzisothiazole; Chlorsulfuron; Corn; Dinoseb; Diuron; Glyphosate; Imazamethabenz; Imazapyr; Imazethapyr; Metabolic profiling; Metabonomics; Naptalam; Neural network; NMR; Quinclorac; Sethoxydim; Sulcotrione; Sulfometuron; *Zea mays*

1. Introduction

The commercial herbicides all act on about 30 biochemically-distinct modes-of-action (MOA), as reviewed by Schmidt (1997). While enzyme assays are available to distinguish these, demonstrating the MOA for a compound is often laborious and time-consuming. In the search for safer and more efficacious pesticides, it is often desirable to: (1) establish which pathway a compound is affecting; (2) determine whether a novel analog has the same MOA as its parent molecule; or (3) classify the MOAs of novel leads found by screening. This should avoid involving well-exploited targets for which novel compounds are not needed (Petroff, 1988; Fiehn et al., 2000; Sauter et al., 1991).

The goal of this paper is to demonstrate that a robust, reliable metabolic profiling method can discern most MOAs targeted by commercial herbicides. We have selected 27 herbicidal compounds representing inhibitors for 19 different MOAs. Plants were treated for 24 h with these compounds and a ^{1}H NMR spectrum of a raw aqueous plant extract was recorded. A computational expert system was developed that can rapidly

* Corresponding authors.
E-mail addresses: karl-heinz.ott@bms.com (K.H. Ott), nelly.aranibar@bms.com (N. Aranibar).
[1] Present address: Bristol-Myers Squibb Co., 311 Pennington-Rocky Hill Rd. 3A-005, Pennington, NJ 08534, USA.
[2] Present address: Bristol-Myers Squibb Co., PO Box 4000, Princeton, NJ 08540, USA.
[3] Present address: BASF Plant Science, 26 Davis Drive, Research Triangle Park, NC 27709, USA.
[4] Present address: 391 South Milton Drive, Yardley, PA 19067, USA.

0031-9422/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S0031-9422(02)00717-3

detect, classify and characterize the nature of the chemical treatment by the changes in the composition of the detected plant metabolites, even under conditions where changes in sample characteristics are very small (often close to the statistical variation between samples).

The term "metabonome" refers to the entire complement of low molecular weight metabolites inside a biological cell, and is also used to describe the observable chemical profile or fingerprint of the metabolites in whole tissue. The metabonome reflects the life history of each individual plant, including age and environmental factors such as soil type and moisture content, temperature, stress factors, and exposure to applied fertilizers and crop protection chemicals. With the expectation that, following exposure to a herbicide, the herbicide's mechanism-of-action might be recognizable in the plant's metabonome, we investigated whether such characteristics can be reliably detected in the NMR spectrum of a plant extract.

The gross chemical composition of various biological fluids has been investigated by a variety of chromatographic and spectroscopic techniques, notably gas and liquid chromatography (Petroff, 1988; Fiehn et al., 2000; Sauter et al., 1991), NMR spectroscopy (Nicholson et al., 1984; Ohsaka et al., 1979; Nicholson and Wilson, 1989; Lee et al., 1991; Bales et al., 1984; Rabenstein et al., 1988; Bell et al., 1987), mass spectrometry (Matsumoto and Kuhara, 1996; Wolfender and Hostettmann, 1996; Aharoni et al., 2002), and infrared spectrophotometry (Jackson and Mantsch, 1996). In animal and human fluids, much of the NMR research has been directed towards disease characterization and diagnosis (Sauter et al., 1991; Nicholson et al., 1984; Ohsaka et al., 1979; Nicholson and Wilson, 1989; Lee et al., 1991; Bales et al., 1984; Rabenstein et al., 1988; Nishijima and Fujiwara, 1997; Somorjai et al., 1996; Holmes et al., 1994; Hahn et al., 1997).

NMR has also provided information on biosynthesis (Lutterbach and Stöckigt, 1995; Prabhu et al., 1996; Weckwerth and Fiehn, 2002), on metabolism (Ratcliffe and Shachar-Hilt, 2001), and on the effects of herbicides on metabolism (Lutterbach and Stöckigt 1994; 1995) and mode-of-action (Hole et al., 2000; Hadfield et al., 2001), or used in investigations of whole plants (Schneider, 1997; Pope et al., 1993). A variety of computational methods have been applied for the statistical analysis of spectral data (Jackson et al., 1999; Shaw et al., 1995; Mansfield et al., 1997; Eysel et al., 1997), including artificial neural networks (NN) (Lisboa et al., 1997, 1998; Anthony et al., 1995; Hiltunen et al., 1995). In many cases, however, it was found that environmental factors contribute significant "noise" to the metabolite profile and reproducibility has often limited the applicability.

Furthermore, in many reports only two states (e.g. normal vs. treated) are simultaneously distinguished. A robust NMR method able to simultaneously detect many different treatment groups has not been described previously. In the search for new pharmaceuticals and crop protection chemicals, it is desirable to have a fast and reliable means to detect the mode-of-action of a new active compound, or pinpoint unusual phenotypes by an altered metabolic profile.

In a recent report (Aranibar et al., 2001), we showed that the ^{1}H NMR spectrum of a crude plant extract provides a fingerprint for the "metabonome", and automated pattern recognition was shown to establish the biochemical mode of action (MOA) for four different herbicide classes. In extension of this earlier work, additional compounds, representing nineteen different MOAs, were selected for simultaneous classification and we present a statistical validation for the methodology.

2. Results

A total of 430 ^{1}H NMR spectra of plant extracts were generated, representing plants treated with four different acetohydroxy synthase (AHAS) inhibitors, four different hydroxyphenylpyruvate dioxygenase (HPPD) inhibitors, two different glutamine biosynthesis inhibitors, and single inhibitors of ACCase, EPSPS, photosystems (PS) I and II, phytoene desaturase (PDS), 4 - hydroxyphenyl - pyruvate - dioxygenase (HPPD), 5-enolpyruvil-shikimate-3-phosphate synthase (EPSPS), glutamine synthase, dihydropteroate synthase (DHP), uncouplers of oxidative phosphorylation, and auxin, as well as systemic inhibitors of microtubule assembly, mitosis/microtubule organization, and cell wall (cellulose) synthesis. Spectra of 80 plants were treated only with the vehicle acetone and represent *controls* in this analysis. Typical spectra are shown in Fig. 1.

One goal of this work is to create a methodology that will enable researchers to rapidly screen novel compounds for herbicidal MOA by comparing their metabolic profile with those of previously characterized standards representing a range of commercially relevant herbicide targets. Model A represents a general-purpose neural network for classification of a wide range of compounds. A second refined model, Model B, is presented that is tailored to distinguish metabolite profiles of treatments that exhibit very small NMR signal differences between each other and/or the *controls*. Both models are cross-validated by using randomly selected subsets for training and testing. The models will be evaluated and applied in simulations to classify compounds novel to the NNs. Lastly, we demonstrate the use of a specialized NN for distinguishing treated from untreated (*control*) plants.

Fig. 2 outlines, in a flow diagram, the procedure used for the analysis presented here. The process reads its input patterns from a database of spectra for all com-

pounds. The user provides a mapping of spectra to user-fined output nodes. We present results for three different levels of output node assignments: (1) compound level: individual compounds each can be assigned a separate class, (2) MOA level: compounds known to affect the same pathway are assigned the same class, (3) treatment level: treated and untreated samples are separated into two classes. A pattern file is then created for all spectra from which a training set, a validation set, and an optional test set are created. The *test set* is created for the *leave-one-out* approach, by selecting a group of patterns corresponding to all spectra of a single compound or a group of compounds. Thus, the test set contains classes (MOAs) or individual compounds that are neither present in the *training set* nor in the *validation set*. The remaining patterns are subsequently divided, by random selection, into two approximately same sized groups of patterns: one used for training (*training set*) and the complementary used for validation (*validation set*). Thereby, each compound's pattern is represented in the *training set* and the *validation set* for *cross-validation*. We iterate over different random selection steps to create a population of 20 NNs. All results presented are averages over such populations. Every time a new test set is generated, the remaining patterns are used to create five new pairs of random subsets. All ten subsets are used to train a NN and classify the pattern present in the test set.

In the following, we will use some abbreviate nomenclature to enhance readability, as follows: For NN classes and associated patterns derived from spectra of extracts of plants that have been treated with a herbicide, we will use the name indicated in column MOA in Table 1 for that herbicide (e.g.: auxin for the pattern representing naptalam-treated plants). If more than one compound is used affecting the same pathway and we want to distinguish the patterns derived from the NMR spectra of the plant extracts individually, we will use the compound generic name, e.g. imazethapyr. "*Controls*" refers to spectra of plants treated only with acetone. *Unknown* refers to a pattern that is characterized by our procedure as unknown, according to the criteria specified in the experimental section. The terms "NMR spectra of plants" (spectra), "patterns for NN analysis" (pattern), and "metabonome" are used interchangeably.

2.1. Model A

Model A encodes one class for *controls* plus 17 classes for the different herbicide MOAs, as listed in Table 1, with all PS inhibitors combined into a single class. Following the procedure outlined in Fig. 2, 20 neural net-

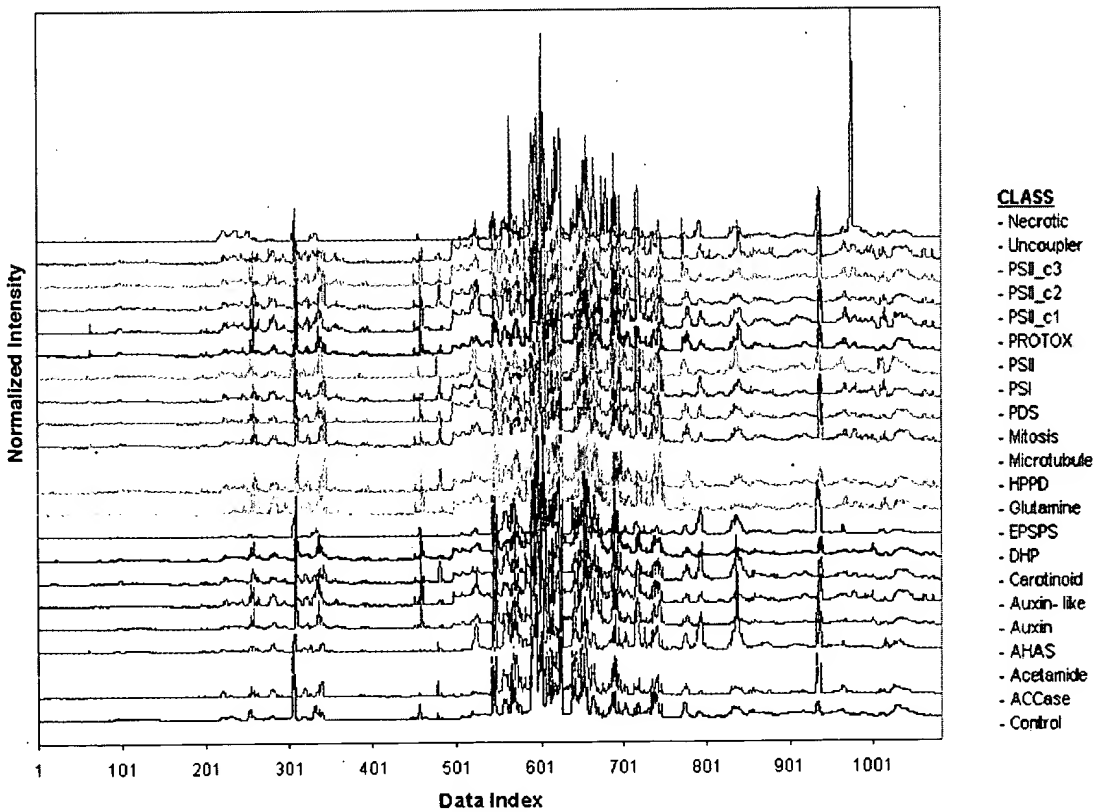


Fig. 1. ^{1}H NMR spectra of plant isolates representing nineteen different MOAs. The spectral region between 9.1–5.7 and 4.5–0.6 ppm is shown and used for analysis. All spectra are scaled to a total mean intensity of 1.0.

Table 1
List of herbicides studied, together with their biochemical mode-of-action and HRAC classification (see Schmidt, 1997)

Class name	Compounds	HRAC class	Mode-of-action
ACCase	Sethoxydim	A	Inhibition of acetyl CoA carboxylase (ACCase)
AHAS	Chlorsulfuron	B	Inhibition of acetohydroxyacid synthase (AHAS, ALS)
	Sulfometuron		
	Imazamethabenz		
	Imazapyr		
	Imazethapyr		
PSII_c1	Lenacil	C1	Inhibition of photosynthesis at photosystem II
PSII_c2	Diuron +	C2	Inhibition of photosynthesis at photosystem II
PSII_c3	Bromoxynil	C3	Inhibition of photosynthesis at photosystem II
PSI	Paraquat	D	Inhibition of photosynthesis at photosystem I
Protox	Acifluorfen	E	Inhibition of protoporphyrinogen oxidase (PPO, PROTOX)
PDS	Norflurazon	F1	Bleaching inhibition at phytoene desaturase (PDS)
HPPD	Sulcotrione	F2	Bleaching inhibition of 4-hydroxyphenyl-pyruvate-dioxygenase (HPPD)
	CL 836057		
	CL 818666		
	CL 836164		
Carotenoid	Amitrole	F3	Carotenoid biosynthesis inhibition (unknown target)
EPSPS	Glyphosate	G	Inhibition of EPSP synthase
Glutamine	Bialaphos[a]	H	Inhibition of glutamine synthase
	Glufosinate		
DHP	Asulam	I	Inhibition of DHP (dihydropteroate synthase)
Microtubule	Oryzalin	K1	Inhibition of microtubule assembly
Mitosis	Propham	K2	Inhibition of mitosis/microtubule organization
Acetochlor	Acetochlor	K3	Acetamide herbicide-like
Uncoupler	Dinoseb	M	Uncouplers of oxidative phosphorylation
Auxin-like	Quinclorac	O	Auxin-like (action like indole acetic acid)
Auxin	Naptalam	P	Inhibition of auxin transport

Class name indicates the name used for the Mode-of-action classes and patterns throughout this paper. CL 836057, CL 818666, and CL 836164 are proprietary herbicide lead compounds of undisclosed structure. + Diuron was applied foliar (class PS II_c2) and systemic [class PS II (root)].
[a] Formulation.

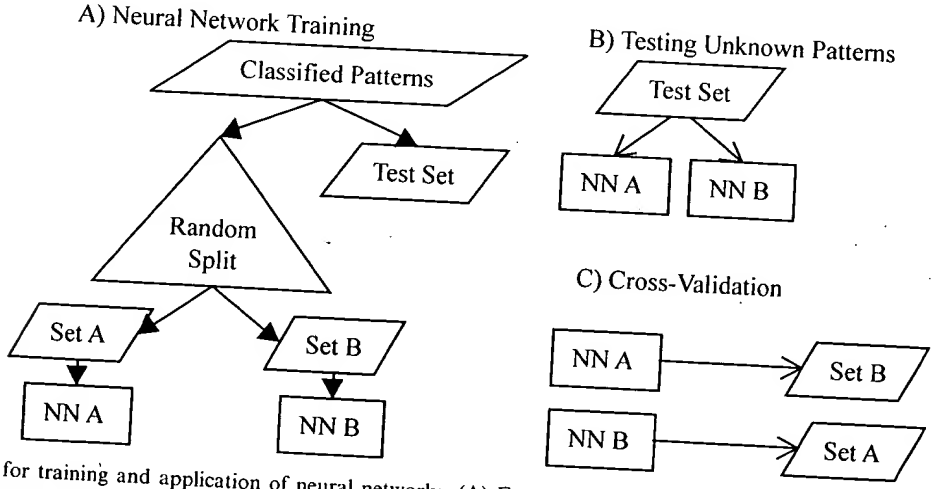


Fig. 2. Flow diagram for training and application of neural networks. (A) From the complete set of all spectra with their associated treatment classifications, two subsets, A and B, of nearly equal size can be created by random selection. Each subset is used independently to train a NN (*training set*), the complementary set is used as the *validation set*. This process is repeated 10–20 times to create a population of NNs. Optionally, specific patterns can be selected into a *test set* for later testing prior to the random selection into subsets. (B) Leave-one-out procedure: the *test set* contains patterns for one or more classes of compounds that are unknown to the NN. The NN is used to classify the *test set*. The pattern of the *test set* should be classified as *unknown* if no other compound was present in the *training set* that represented the MOA of the pattern in the test set. (C) Cross validation: the *validation set* is classified by a NN to produce statistics on the sensitivity and selectivity of the classification for compounds with pathways that are known to the NNs.

works were trained with randomly chosen subsets of the available spectra, and the complementary set of patterns were classified by the NNs. The results are summarized in Fig. 3, which shows graphically the average number of *correct*, *wrong*, and *unknown* classifications of the spectra of the *validation set* by the 20 different NNs. Overall, 64% of the spectra were classified correctly on an individual basis, and 30% of the spectra were classified as *unknown*.

Inhibitors of pathways affecting amino acid pools (e.g. AHAS, EPSPS, glutamine biosynthesis), fatty acid synthesis (ACCase) are consistently recognized, as is the photosystem II inhibitor, diuron when applied to roots. With only 6% of the samples classified as *wrong*, there is little confusion between the different classes, and most wrong assignments are observed in only one of the twenty different NNs. Some wrong assignments are observed between related MOAs. An unusually large fraction (10%) of glyphosate patterns is confused with AHAS inhibitors (discussed below). Other patterns, such as PROTOX, DHP, and, most notably, patterns of herbicides affecting the auxin transport, microtubule formation and mitosis have an increased pool of *unknowns*.

Confusion with *controls* is observed for several treatments in a few isolated cases (1–5%), but only Auxin patterns have significant percentage (20%) confusions with *controls*. Inspection of the NMR spectra reveals that many treatment pattern, most notably Auxin, Microtubule, and Mitosis show very little difference between each another and to the *control* samples. The microtubule inhibitor treated samples are also confused with HPPD inhibitors (7% wrong). Separate analysis shows that this confusion is largely caused by the inclusion of two very weakly herbicidal compounds into the HPPD class. The photosystem inhibitor class is assigned to several inhibitors that have, in turn, large fractions of *unknowns*. A similar calculation representing four separate PS classes for a total of 23 different classes, produces almost identical overall results (62% correct/ 27% *unknown*), and only small changes in the confusion between the different classes.

2.2. Model B

After identification of several batches of treatments by the NN described in Model A, we removed those treatments groups and performed a second round of classification for the remaining MOAs that had more than one third *unknown* classifications and were found to be more likely to be confused with one another. We also refined the analysis by using separate classes, PS I and PS II c1, c2, and c3, for the photosystem inhibitors. The refined NN (Fig. 4) improves the classification by removing some over-represented and strongly distinct signals, to focus on smaller differences between the remaining patterns. Overall, the recognition level has risen by about 20% for the MOAs that had previously been difficult to classify. In particular, microtubule, mitosis, and auxin are now more often recognized.

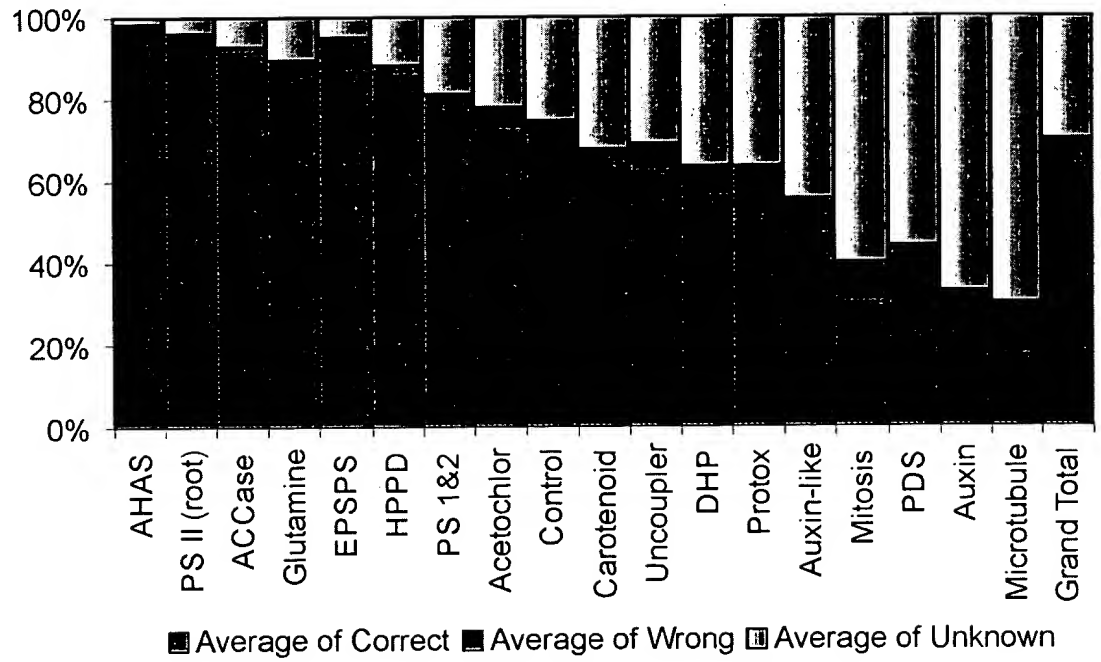


Fig. 3. Average number of *correct*, *wrong*, and *unknown* classifications of the NMR spectra by 20 different neural networks in Model A. A randomly selected subset of ca. half of the spectra was used for training, whereas the complementary set (not used in training) was classified automatically by the trained neural network. PS 1&2 refers to a class that is trained with all photosystem inhibitors.

The confusion matrix (Table 2) indicates that, while there are more frequently classifications confused between related pathways, PS I and the three different subclasses of PS II inhibitors separate well. PS II c3 has a metabolite profile that is very distinct from that of the other PS inhibitors while PS I, PS II c1, and PS II c2 have more closely related profiles. For example, about 10% of PS I pattern are classified as PS II c1 in average over all simulations. Similarly, 12% PSII c2 inhibitors are classified as PCII c1. Thus, the second step which is introduced in an attempt to enhance the sensitivity of the approach, simultaneously enhances selectivity.

Auxin and DHP get confused in some of the runs, which is reflected in increased percentage of wrong classifications for these classes, and also in a higher fraction of *unknown* classifications. Again, we do find

Table 2
Confusion matrix for Model B

Model B	Classification as percent recognition													
Actual class	PDS	PROTOX	PSII_c1	PSII_c2	PSII_c3	PS I	Uncoupler	Auxin-like	Auxin	DHP	Microtubule	Mitosis	Acetochlor	Unknown
PDS[a]	31					1			3	2				58
PROTOX		80	1								6	2		11
PSII_c1	3	1	50	3	1	1				1		2		38
PSII_c2			12	67								1		21
PSII_c3					93									7
PS I			10		1	53					1			35
Uncoupler				1	1		80		1		1			16
Auxin-like								71	1		1			27
Auxin	1								51	7	2			39
DHP									1	68	1			30
Microtubule		2						5	3	3	51	5		32
Mitosis	1		1	4							3	50	1	40
Acetochlor								1		3			70	21

Only MOAs were presented for which Model A lacked sensitivity. Rows indicate the actual treatment and columns represent the averages for the assignment by 20 independent NNs. The diagonal elements of the confusion matrix represent percent correct assignment whereas (non-zero) off-diagonal elements imply confusion between classes.

[a] PDS is represented by only six spectra in total.

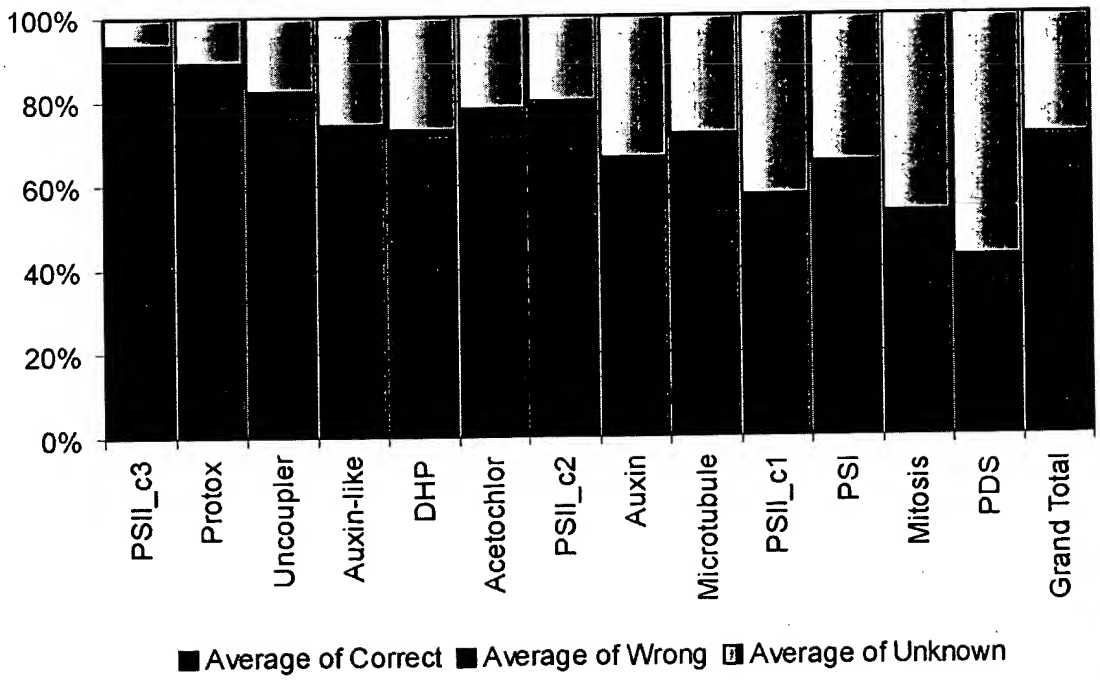


Fig. 4. Summary of classification results for Model B. Averages of percent recognition of total for each compound/MOA is shown for 20 NNs, each trained with a different random selection of half the spectra classifying the complementary set. The compounds tested were all represented in, but not part of, the training set. Each column corresponds to a different class.

that microtubule and mitosis have a "weak" signature that is frequently, but not consistently, confused with other MOAs.

The results of the two calculations listed above are not only representative by virtue of performing repeated runs with different selections of spectra used to train the network. We find that very similar results are achieved when classification schemes are changed. The best overall results achieved so far with this data set are for a 15-class NN (classes: *control*, AHAS, HPPD, PS II (root), glutamine, PSI, PS II c3, EPSPS, carotenoid, protox, PS II c1/c2, auxin-like, DHP, uncoupler, acetochlor) where PS II c1, and PS II c2 are combined into a single class, and microtubule, mitosis, and auxin inhibitors are not part of the training. This NN has overall 85% correct, with > 70% recognition for any included MOA, and only 13% *unknown* and 2% *wrong* classifications.

2.3. Application of Models A and B

How do the models presented in calculation A and B perform when a new compound is presented that is not part of the training set? Which MOAs are easily confused with others? How sensitive and how selective is the method in situations with overlapping or partially divergent MOAs?

To answer these questions, we designed a leave-one-out procedure in which we remove one compound at a time from the data set and calculate 10 NNs, using 10 different random selections of half of the remaining spectra for training (the other half is disregarded). We then present the pattern removed in the beginning to the NN for classification. If a compound is novel to the NN and there is no related compound in the training set, we expect the NN to issue an *unknown* classification. If other compounds representing the MOA of the compound presented are in the training set, we hope to find this compound to be correctly classified. Related MOAs are expected to be partially activated. Partial activation is represented in the NN in the actual activation values of the output nodes. Since those numbers are difficult to present in the format of a publication, we use the average of *correct* classifications over a series of related networks as a measure of relatedness, given the rules laid out in the experimental section.

The results of the leave-one-out procedure are summarized in Figs. 5 and 6. We will discuss four different situations: (1) a group of chemically diverse compounds has the same MOA; (2) a group of compounds from a series believed to target the same enzyme are metabolized differently by the plants; (3) A group of com-

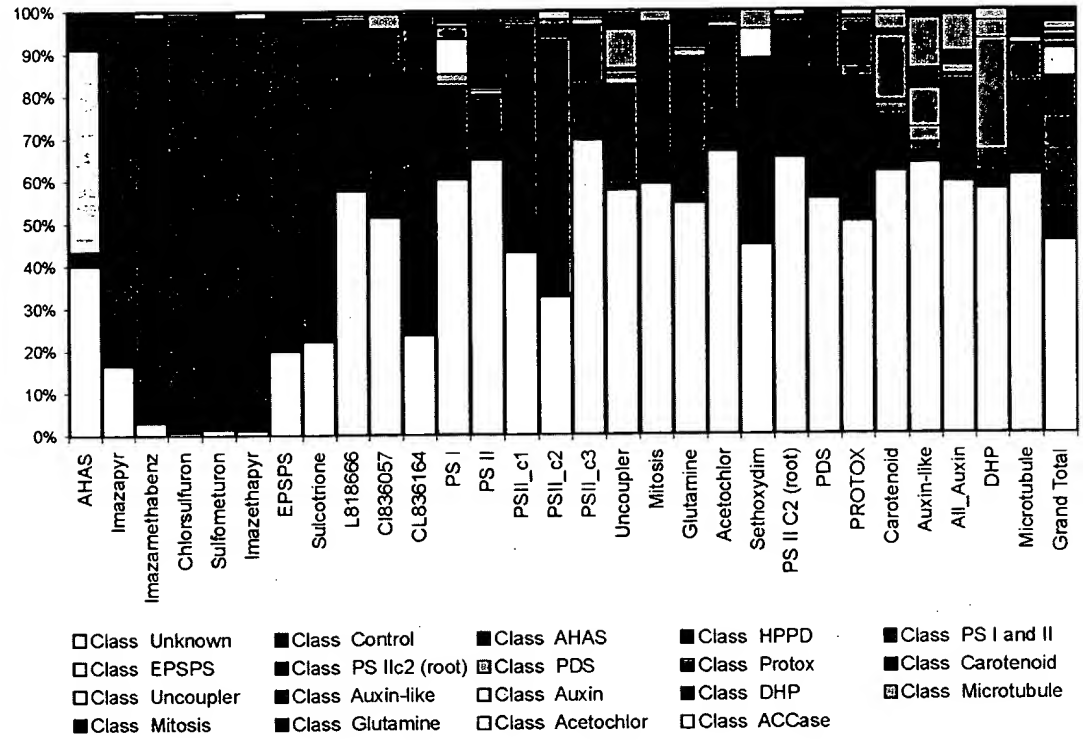


Fig. 5. Results from "leave-one-out" computations. Each bar represents average classification result of 10 NNs for the compound/compound group indicated. Each group of 10 NNs was trained with all spectra except those for compounds or groups of compounds indicated on the horizontal axis. The colors refer to the class the classified spectra were assigned to. For example, in the first bar, all AHAS inhibitors were removed before training 10 NNs with randomly selected subsets of 50% of the remaining patterns. The AHAS inhibitors are classified as ~40% *unknown*, 48% EPSPS, 8% carotenoid, 5% *control*.

pounds affects different steps in the same pathway; (4) A compound represents an entirely new MOA.

2.4. Co-classification of chemically distinct compounds by their common MOA

The imidazolenone and the sulfonylurea herbicides, as well as many other commercial herbicides, inhibit the AHAS enzyme. We chose five of these herbicides having a range of different specificities, but all targeting AHAS. We had previously shown that a NN trained to recognize the metabonomes of plants treated with imazethapyr, glyphosate, two other herbicides, and controls recognizes >99% of the metabolite profiles of other AHAS inhibitors into the AHAS MOA. Extending this approach, we now included more MOAs into the NN models and performed a more rigorous, cross-validated approach.

Removing one compound from the training set, leaving four compounds as AHAS representatives for training, more than 90% of the samples are classified correctly, with most AHAS inhibitors having more than 95% *correct* classifications. Imazapyr has only 83% *correct* classifications and 17% *unknown* classifications. This result reaffirms our earlier findings (Aranibar et al., 2001), but now, the statistical significance is higher since the recognition is above the background of many more alternative MOAs.

Using only one of the four AHAS inhibitors together with all other MOAs in the training of the NN, decreases the sensitivity as there are only about six compounds remaining in the training, resulting in about 20–30% *unknown* classifications. However, of the positive classifications, ~80% are true positive assignments. This average is reduced by over 10% by poor recognition when imazethapyr is used as representative for the AHAS MOA within the training set. We attribute this to the divergence between the individual NMR spectra since the imazethapyr samples had been collected in the very beginning of the study when we lacked experience in reproducibly collecting the samples, and the growth chamber was set 3 °C lower. Most of the difficulty in recognizing different compounds affecting the AHAS enzyme are caused by the presence of glyphosate as a EPSPS inhibitor with a similar metabolite profile.

If all AHAS inhibitors are removed from the training, AHAS becomes a novel MOA for the network. In this case we find that about half the samples treated with AHAS inhibitors are (wrongly) classified as EPSPS inhibitors, and about 40% are *unknown*, as expected. Also, vice versa, glyphosate will be classified as an AHAS inhibitor if no sample from a glyphosate-treated plant was present during training. AHAS and EPSPS are in different pathways, and in general, the network is capable of separating these MOAs, as long as the NN

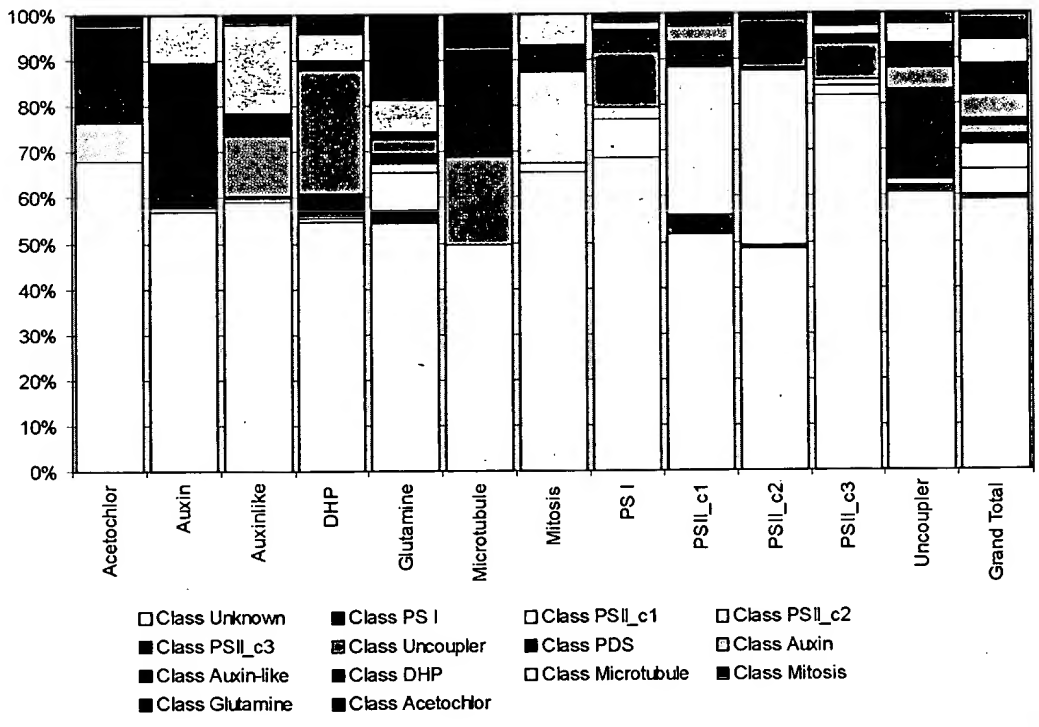


Fig. 6. Classification of compounds unknown to the NN (i.e. not included in training). Each bar represents the classification of the compound indicated on the horizontal axis by 10 different NNs. The NNs were untrained in the compound/MOA presented, but trained with all other MOA. The compounds presented to the NN represent a "Novel MOA" to the NN. An *unknown* classification is the expected correct answer.

has been trained to do so. However, there is an average of ~10% of glyphosate samples assigned to AHAS even when glyphosate is represented in the training. (The higher variability in the imazethapyr NMR spectra discussed above, is mostly responsible for the false positive assignments.) The NMR spectra of AHAS and glyphosate treated plants are very similar with only very few proton resonances different between the two populations while there are a considerable number of signals that commonly change with respect to the *control* and other MOA spectra. Many of those signals can be assigned to amino acids and we find that inhibition of amino acid metabolism can increase the pool of free amino acids, presumably due to increased protein turnaround. While the composition of the amino acids found changed in both populations is different, the communality dominates if the NN is not specifically trained to recognize the smaller differences. Thus, both MOAs share similarities in the resulting metabolite profile. The differences are due to the levels and types of amino acids that accumulate.

Inhibition of glutamine synthase, in contrast, has a very different profile, lacking the increase of amino acid pools but distinguished readily by several resonances and we attribute several of the resonances of the glutamine biosynthesis inhibitors to components of the formulation rather than to natural metabolites.

2.5. Same target, different metabolic fate

As a challenging example relating to a lead optimization problem, we had selected three chemically analogous compounds from a series of experimental HPPD inhibitors, and sulcotrione as a commercial herbicide representing a different chemical class. Corn is resistant to sulcotrione. From the remaining compounds, one compound is highly active, one is very weakly active in vivo, but was predicted as highly active in a quantitative structure–activity relation (QSAR) study (data not shown). The last sample appears much more potent than was predicted by QSAR. Since this set is so diverse in its in vivo activity, the signatures are less distinct in the context of the many other MOAs. This is reflected in an increased number of *unknown* classifications, ranging from ~25 to ~55%. However, the correct MOA assignment still dominates the positive classifications in all cases and a more specialized NN can also highlight the more subtle differences between these compounds. When using each compound, in turn, as representative in the training, the very active compounds reveal a very similar profile, while spectra of the very weakly herbicidal compound are often confused with *controls*, and patterns that are very similar to those of *controls*, like Microtubule. (Removal of the weak HPPD inhibitors from the training set does, in turn, improves slightly the sensitivity of recognizing of some of these patterns.)

2.6. Pathway recognition

Co-classification of PS inhibitors into a single class is a model for recognizing compounds that inhibit different related biochemical functions. Fig. 5 demonstrates that the photosynthesis inhibitors do, to some degree, co-classify if the network is trained with a combination of three of four of the PS I, PS II c1, c2, c3 inhibitors. PS II c1 and PS II c2 are well recognized into a related class with most of the positive classifications being correct. The results (~1/2 *unknown*, 1/2 shared class assignments) are similar to the pattern observed for the HPPD inhibitors as described above. The majority of the positive classifications of PS 1 are also correct, but several other MOAs have a similar large percentage (20–30%) classified as photosynthesis inhibitors.

Applying the more specific and refined model B that has each PS inhibitor as a separate class (Fig. 6) indicates, in concordance with the analysis of the confusion matrices during the validation runs, that while PS II c1 and PS II c2 have closely related profiles, PS I is more distinct. PS II c3 has little in common with the other PS inhibitors, but shares some features with uncouplers.

2.7. Novel MOAs

Several of the MOAs are represented by a single compound in the present study. Thus, removing these compounds before training the NN simulates results for compounds belonging to novel MOAs. We would desire that compounds belonging to a MOA that was not represented in the training should be classified as *unknown*. Every other classification would be considered *wrong*. For many compounds presented to an NN, we find that, for new MOAs, about 60% of the classifications are in fact *unknown*. The remaining 40% are variable classifications. For practical purposes, we are mostly concerned when a single "*wrong*" classification dominates, since this could cause false positive conclusions. Using Model A, several compounds have 20–30% of their patterns classified incorrectly as *control*. Application of the control model (below) can characterize these compounds as "*treated*" and thus identify them as novel MOAs. Incorrect classifications as *controls* appear to be an indication that there is very little change in the metabolite profile caused by these compounds. Those changes will only be picked up if such a MOA is specifically presented to the NN. In addition, the NN training over-weights the untreated samples (e.g., 80 *controls* vs. 12 *treated* spectra), and the *controls* show greater experimental variation due to our experimental design.

Applying our more specialized NN, Model B, also overcomes many of false positive classifications, as illustrated in Fig. 6. Now, all patterns have more than 60% *unknown* classifications, one of our empirical cutoffs for novel MOAs. The majority of the "novel compounds"

has no consistent *wrong* classifications to another class and can be attributed to noise, i.e. experimental variability, especially for treatments that cause little change in the metabolic profile.

Auxin and DHP have about 24% classifications confused between each another. We found, by comparison of the NMR spectra, that one batch of DHP has a very distinct metabolite profile from that of auxin, but the other batch of DHP lacks several metabolites present in the first batch and resemble more closely spectra of *control* and auxin.

2.8. Control model

Specialized NNs that are optimized to recognize a specific treatment versus all others can be more sensitive and specific. From the results presented above, it is apparent that distinction of samples treated with a compound versus samples treated with a blank solution sometimes poses difficulty. In the following we evaluate whether a specialized NN to distinguish treated and untreated samples might further reduce the already small error rate. In a modified calculation, the samples were classified into two subsets, *treated* and *control*, i.e. those treated with a compound solution and those treated with a blank solution. We calculated the average over ten classifications using the cross-validation procedure outlined in Fig. 2.

As shown in Table 3, 96% of the spectra from treated plants are recognized as *treated*, and only 3% were false negatives, if other spectra of the same treatment were included in the NN training. *Controls* are recognized as such in 82% of all cases, with 15% false positives (*controls* misclassified as *treated*).

To further validate the control model using the leave-one-out method, we also removed, in turn, one compound, and also all AHAS, and all HPPD inhibitors at once, to simulate how such a binary model would perform when a new compound, previously unknown to the network, would be introduced. If a particular treatment was not known to the NN, the average true positive rate for the data is still 89%, with 9% false negatives, as shown in Fig. 7, indicating that there is a strong signature that characterized treated plants.

As expected, best results are usually achieved if other compounds of a series, or with a similar MOA are included in the training. Most HPPD and AHAS inhibitors are consistently classified as treated, as long as other inhibitors of that class are included. Even if all AHAS pattern are excluded from the NN training, the patterns are still recognized as *treated* in >95% of all cases. Many other inhibitor patterns, like patterns of photosystem inhibitors, are also well recognized, possibly due to partial overlap with patterns included in the training.

Table 3
Statistics for the control model

Control model	Classification as percent recognition		
Actual class	*Treated*	*Control*	*Unknown*
Treated (*known*)	96	3	1
Treated (*unknown*)	89	9	2
Control	15	82	3

Treated (*known*) refers to average results of 10-fold cross-validated NN runs in which all MOAs were part of the training procedures. *Treated* (*unknown*) refers to the results of runs in which, in turn, each compound or MOA group was first removed from the data set, after which the 10-fold cross-validation procedure was run, and the spectra of the compounds/MOAs that were excluded were classified by the resulting NNs. This simulates the NN classification for a novel compound or new MOA.

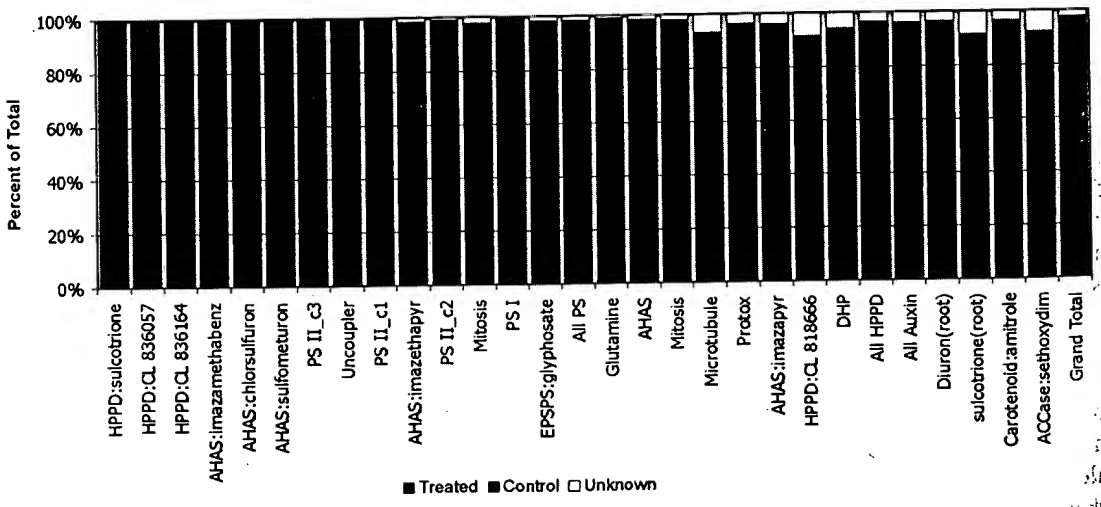


Fig. 7. Results of recognition of unknown treatment classes by the control model. Each column represents the average fraction of *correct*, *wrong* or *unknown* classification of 10 NNs that were trained without the compound or group of compounds used in the training.

3. Discussion

3.1. Growing conditions

One of the most important requisites for the work on metabolic profiling in plants is the stability and reproducibility of the physical conditions in which the plants are grown. Plants, as all living organisms, react to different environmental stimuli and changes that turn on and off different genes expressing different proteins and enzymes, and developing different metabolic states, usually the most appropriate for the best development of the organism in the given environment.

In the early developmental stage (5–10 days after germination) in which the seedlings in this study were treated and harvested, metabolic changes are fast and changes in the concentrations of metabolites are considerable for the small amount of growing point tissue that can be collected. Relative small changes in the environment of a plant can be reflected in very detectable variations in the absolute concentration of a metabolite and with that, a change of the profile.

For these reasons, the use of growing chambers, where the environmental conditions can be accurately controlled, is mandatory. In the course of the present study, for example, some plants had to be transferred from one growing chamber to another, due to the mechanical failure of the first one. Several hours at a more elevated temperature and then change in illumination produced detectable differences in the metabolic profiles. The NN can be trained to either recognize or ignore these changes in environmental conditions. Thus, it is clear that the use of green houses and field plots are not appropriate for growing the plants used in this kind of study. This observation may have implications for other kinds of profiling, e.g., gene expression profiling.

3.2. NMR spectroscopy

The use of an acidic matrix to prepare the extracts of plant tissue allowed us to isolate the widest range of primary metabolites (amino acids, sugar, sugar-alcohols, organic acids, etc.). Due to the relative low sensitivity of NMR spectroscopy, it is important to choose as many of the metabolites present in the highest concentrations as probes for the total metabolic profile. This extraction matrix does not produce any undesirable solvent peaks in the NMR spectrum. Reproducibility of the NMR operating conditions is the key for a reliable classification of the spectra. Temperature and spectral width seem to be the most important factors. The exact total concentration of metabolites in the sample (which is dependent on the amount of tissue used for extraction) is less critical for two reasons: (1) use of an internal reference standard in each sample, and (2) normalization of all the spectral intensities as part of the pre-processing of the spectra when preparing patterns for analysis with the neural network software.

Many replicates of each sample were prepared and measured in each experiment. Usually 5–12 plants were grown, treated, and harvested for each treatment class. Each experiment was repeated at least twice at different times. We find that there is, even under tightly controlled condition a slight "batch" factor in which samples of one batch tend to cluster together. This only becomes a problem when experimental conditions have changed or if the discrimination is already weakened by other factors, such as too many similar pathways spread over too many nodes. Since NNs can be trained to recognize fluctuations in conditions, it is recommended to always include, with each batch of treatments to be classified, a few reference samples of the MOAs that are most likely to be targeted by the compounds under investigation.

3.3. Pattern recognition

We have presented the results for a full NN model that simultaneously recognizes a wide variety of metabolic profiles with a high success rate and confidence. Most importantly, we find that compounds affecting the same MOA have related NMR spectra and can be distinguished from a wide range of other MOAs with high confidence. Compounds not previously known to the NN co-classify with other compounds affecting the same MOA. Related MOAs are sometimes indicated by an increased fraction of patterns of a treatment being classified to a second MOA.

MOA classes that are part of the NN *training set* are usually well recognized. Inhibitors that affect pathways that are involved in the metabolism of common, soluble cellular components, for example inhibitors of the amino acid metabolism pathways, are the most distinct and are detected with high confidence. Other inhibitors do not create large changes in the profile of soluble compounds compared to *controls*: the auxin, mitosis, and microtubule MOAs are difficult to classify in the background of the many other compounds and produce a larger fraction of *unknown* classifications. Nevertheless, even in these more difficult cases, there are only a small fraction of false positive classifications and even those samples are classified with high confidence by the control model as *treated*.

The NN method is often capable of handling closely related pathways, and we find that the analysis of the confusion matrix for compounds affecting closely related MOAs yields fruitful insights in the particulars of each compound, and highlight similarities as well as differences in their activity and metabolic fate. For example, we found confusion between patterns of PS I, PS II c1, and PS II c2, but not between these patterns

and that of PS II c3. Thus, the separation and analysis of the confusion pattern yields insight into the response patterns that are created by the different inhibitors of the photosystem I and II and their subsystems. The analysis of the confusion matrix for NNs trained with a single inhibitor of a series, classifying other compounds in a series, as discussed for the HPPD inhibitors yields deeper insight into the differences in metabolic fate. Compounds not active in corn due to their limited uptake or rapid metabolism co-classify with highly herbicidal compounds of the same chemical family, but at a reduced NN output activation level. In addition, the alteration in the metabolic fate may also be indicated when samples of a treatment are also classified by the NNs into other classes at elevated percentages (>5%). For novel compounds or compounds for which the MOA is not well established, the MOA might not be represented in the *training set*. We simulated this scenario by removing a complete class of compounds prior to training. The results of the "leave-one-out" experiments highlight a critical feature of the method. A NN trained to discern *treated* and untreated samples classifies active herbicides with negligible small false negative rate to the treated group (see discussion of the Control model). In a detailed model, like Model A or B, novel compounds are generally assigned to the correct MOA or pathway if this pathway has been defined during the training by the NN. Furthermore, if the pathway is not known to the network, that is the NN has not seen a mechanistically-related compound, we are likely to get a majority of *unknown* classifications. If related pathways are present in the training, we are likely to find that more than 20% of all classifications point to the related MOA(s). We find such a situation for the related PS inhibitors and the HPPD inhibitors that have very different activity levels. Compounds of sufficient high herbicidal activity affecting the same MOA co-classify at a high proportion. However, caution is indicated when a novel compound affects an MOA that is not known to the NN and the profile of the novel MOA has many overlapping features with an MOA that is known to the NN (the confusion of AHAS inhibitors and glyphosate demonstrates this). The best safeguard against this type of "false positive" is the inclusion of as many MOAs as possible into the NN and the observation of additional experimental evidence, e.g. the plant phenotype.

The general purpose model, Model A, produces satisfactory results for many MOAs and might suffice in praxis for many applications. The model can be generalized in many ways, and other class assignments can be chosen. In the variations we studied, we found little change in the overall success rate upon using different classification schemes (like various combinations of MOAs in single or split classes), as long as treatment classes were not entirely removed. The particular models detailed in this report were chosen to exemplify different levels of refinement and a stepwise approach that is most likely to be used in a research setting.

The two step procedure was guided by our quality control procedures that had indicated that there are spectra, that include photosystem, mitosis, microtubules, auxin classes, etc., that are statistically very similar (data not shown), an observation that is confirmed by visual inspection of the overlay of the NMR spectra. Also, *controls*, AHAS and HPPD inhibitors were largely overweight in the training of Model A, since multiple compounds of the same MOA were present. Model B has the treatment regimes more equally represented.

Because these experiments are subject to normal biological variation, it is unrealistic to expect 100% accurate classification at all times. Some plants might be less susceptible to a given herbicide than others and their metabonome would be less affected. In such cases, a treated plant might be wrongly classified as a *control*. Extraneous effects might cause changes in the NMR spectrum, causing classification of some treatments as *unknown*. Different MOAs that result in similar metabolite profiles will be confused with each other, while other MOAs might have too small an effect on the NMR spectrum to be classified. Ultimately, it will be necessary to set a threshold or cut-off for acceptance of a correctly classified MOA. In general, we find that if more than 80% of all patterns of a batch are classified consistently, these assignments can be trusted. As the *unknown* fraction for a batch approaches 50%, increased caution is advised.

In cases where the NN responds to new samples with over 60% *unknown* classifications, the MOA of the new sample might indeed be *unknown*. A specialized NN including only those MOAs that have similar metabonomes can improve selectivity. Those compounds that retain a large number of *unknown* classifications and also have a larger number of confusions with other MOAs or *controls* will need close scrutiny.

The particular choice of class assignments, classes included in training, the mix of spectra included in the training and other factors seem to affect the particular outcomes to less than 3%, on average. This implies that the operator has only very limited influence on producing a particular outcome, except for avoiding particular MOAs or compounds.

Most variability between the NMR spectra within a group is found for *controls*. Every batch was accompanied by controls, leading to many more samples, and thus reflecting the overall variation between the batches over a period of more than 1 year. Also, plants that did not grow to the required size before treatment were sometimes included as *controls*. Most notably, we observe that the reproducibility between all spectra has increased with the experience of the scientist running the studies such that for the first few batches, correlation

coefficients between samples (regardless of treatment) is better than 0.8, while after all details of the procedures were carefully established correlation coefficients were consistently better than 0.9.

Most false negative (*control*) assignments are attributed to the lack of a sufficiently diverse training set. For example, only a few MOAs are represented by compounds applied to the medium and thus our control model lacks mostly in recognizing pattern for compounds applied though the medium. The same compounds applied foliar are recognized as *treated*. A more representative data set, with more compounds for each class, and consistent application schemes should overcome these difficulties.

Selectivity and sensitivity depend to a large degree on factors other than the patterns themselves, for example the presence of MOAs used in the training and the granularity of the class assignments for related MOAs. Manipulation of the analysis scheme can achieve increased success rates. If the operator has some knowledge of the MOAs that a set of compounds might affect, it might be advisable to reduce the number of MOAs in the *training set*. More specialized NNs often will show increased robustness of the assignments. Conversely, selectivity and sensitivity drop when the NN is forced to separate between signatures for closely related patterns, i.e. to distinguish too many closely related pathways. However, we strongly favor inclusion of several MOAs in the *training set* to avoid creating signatures that are unrelated to treatment per se, such as stress markers, rather than a specific compound profile. In particular, "false positive" assignment (assignment of a compound to the wrong MOA) can largely be avoided when enough related MOAs are included to act as positive controls.

4. Conclusions

This work has shown the feasibility of ^{1}H NMR spectroscopy of plant extracts, in combination with artificial neural network analysis, to distinguish treated from untreated (*control*) samples and discriminate, with high reliability, the modes-of-action of many different, commercially important herbicides. Easily obtainable extracts from plants, analyzed by 1D ^{1}H NMR contain a wealth of information about the treatment of the plants. NMR is sensitive enough to produce fingerprint information that enables the researcher to discern between related MOAs and about twenty MOA classes have been discerned by the automated pattern recognition approach. Compounds affecting the same target enzyme are classified by their metabolic profile to the corresponding MOA, even if only one reference compound is used to create the signature for that MOA. Compounds with novel MOAs are classified as *unknown*. Detailed analysis also highlights differences between compounds of a series that affect the same target but that are being metabolized differently. Of the 19 MOAs studied, the control group (untreated), AHAS, HPPD, ACCase, EPSPS, PROTOX, carotenoid, PS-I, uncoupler, auxin-like, acetochlor, PS II, and glutamine synthase inhibitors were all well classified (little or no confusion with control plants or other MOAs). For MOAs that have closely related metabolite profiles, enhanced sensitivity is achieved when a specialized NN is used that includes only the closely related MOAs. Such a stepwise process can be included into an expert system to classify metabonome profiles of all treated plants with high confidence. The method is reliable when the experimental conditions are well controlled and accurately kept under standard conditions. There exists a large potential for similar applications in the agricultural and pharmaceutical industries, as many biological tissues are amenable to study by metabolic profiling.

5. Experimental

The plant preparation methods were as described previously in Aranibar et al. (2001). In brief, *Zea mays* seeds (Pioneer 3514) were set to germinate for 5-days in a controlled-environment growing chamber. The plants were treated post-emergence with the herbicides shown in Table 1. Twenty-four hours post-treatment, the plants were harvested and the meristematic tissue (approximately 250–300 mg per plant) was collected, flash frozen in liquid nitrogen, and stored in a liquid nitrogen freezer until further use. The plant meristems were then pulverized, suspended in 0.25 N HCl, and centrifuged. The supernatants or plant isolates containing the soluble metabolites were separated and reserved for ^{1}H NMR spectroscopy.

For each compound, treatment was repeated in at least two separate batches, each containing six individual plants, resulting in at least 12 spectra per compound. While conditions were kept as constant as possible for the treated plants, some of the control plants reflect small variations in environmental conditions and growth stage. The batches of plants were spread over a period of more than 1 year, and a few plants were grown at a slightly elevated temperature (due to a malfunctioning temperature controller). Treatment of plants with AHAS inhibitors, sethoxydim, glyphosate, and two batches of diuron were applied to the media, while all other inhibitors were applied to the leaves ("foliar"). The following data were excluded from most of the analysis due to the lack of sufficient samples for randomized training and testing: (1) two glyphosate treated plants were killed rapidly and were decaying after 24 h; (2) a single batch of six PDS treated

plants was ignored in some analysis, due to the lack of a second batch; (3) for the detailed analysis in the latter part of this paper, we removed one batch with six samples of *control* plants and twelve samples of imazethapyr-treated plants because the NMR spectra were recorded at a higher temperature; (4) we also removed one *control* sample that showed strong stress response signals.

The NMR profiles were classified using a supervised pattern recognition approach in which a neural network is "trained" using a set of NMR spectra for plant extracts whose origin and nature is well known, i.e. with known herbicide treatments, known genetic phenotypes, etc. The NMR spectra are "memorized" as patterns during the neural network training step. When the spectrum of an "unknown" extract is presented to the trained network, it will be recognized only if it is a member of the training set; otherwise, it will not be recognized and will be flagged accordingly.

The SNNS (Stuttgart Neural Network Simulator, University of Stuttgart, Stuttgart, Germany) software was encapsulated into a user interface that reads as input a definition of a network topology, spectra to be used to train the network, and spectra to be classified. The output of the classification run is analyzed automatically and converted into tabular and graphical form. For the NN, a three-layered, fully-connected topology is defined with 1080 input nodes (representing the spectral data points after preprocessing), 12 hidden nodes, and up to 30 output nodes. All nodes are characterized by a logarithmic input function and unity output function. Random values are assigned to each parameter initially, and the resilient backpropagation algorithm is used for optimizing the weights, which are updated for 500 iterations in topological order. We use an initial update value of 0.1 and a maximum step size of 50. The NN is trained by presenting a subset of the pattern to a suitable network topology and, after training, the network can classify the metabonome represented by the NMR spectra of samples other than those used in the training. The output of the classification step is in the form of output unit activation values. The procedure employed converts the activation values into a more readable classification by assigning a classification to the spectra if a single output node has an activation value >0.6 and no other output node has activation values >0.4. Otherwise, the spectrum is classified as *unknown*. The classification for each spectrum by the NN is recorded and compared to the actual treatment of the corresponding plant. The number of *correct* and *wrong* classifications are tabulated, and are shown as bar-graphs, together with the spectra that were classified as *unknown* by the NN and that are counted separately. The classifications are also displayed in the form of a Confusion Matrix, whose rows indicate the actual treatment and columns represent the assignment generated by the NNs. The diagonal elements of the confusion matrix represent correct assignments, whereas (non-zero) off-diagonal elements imply confusion between classes. In addition, analysis can be performed for batches of samples that received the same treatment rather than an individual sample, thus reducing the possibility of false conclusions.

References

Aharoni, A., De Vos, C.H.R., Verhoeven, H.A., Maliepaard, C.A., Kruppa, G., Bino, R., Goodenowe, D.B., 2002. Nontargeted metabolome analysis by use of Fourier transform ion cyclotron mass spectrometry. OMICS 6, 217–234.

Anthony, M.L., Rose, V.S., Nicholson, J.K., Lindon, J.C., 1995. Classification of toxin-induced changes in ^{1}H NMR spectra of urine using an artificial neural network. J. Pharm. Biomed. Anal. 13, 205–211.

Aranibar, N., Singh, B.K., Stockton, G.W., Ott, K.-H., 2001. Automated mode-of-action detection by metabolic profiling. Biochem. Biophys. Res. Commun. 286, 150–155.

Bales, J.R., Higham, M., Howe, I., Nicholson, J.K., Sadler, P.J., 1984. Use of high resolution nuclear magnetic resonance spectroscopy for rapid multi-component analysis of urine. Clin. Chem. 30, 426–432.

Bell, J.D., Brown, J.C.C., Nicholson, J.K., Sadler, P.J., 1987. Assignment of resonance for acute phase glycoproteins in high resolution proton NMR spectra of human blood. FEBS Lett. 215, 311–315.

Eysel, H.H., Jackson, M., Nikulin, A., Somorjai, R.L., Thomson, G.T.D., Mantsch, H.H., 1997. A novel diagnostic test for arthritis: multivariate analysis of infrared spectra of synovial fluid. Biospectroscopy 3, 161–167.

Fiehn, O., Kopka, J., Doermann, P., Altmann, T., Trethewey, R.N., Willmitzer, L., 2000. Metabolite profiling for plant functional genomics. Nat. Biotechnol. 18, 1157–1161.

Hahn, P., Smith, I.C.P., Leboldus, L., Littman, C., 1997. The classification of benign and malignant human prostate tissue by multivariate analysis of ^{1}H magnetic resonance spectra. Cancer Res. 57, 3398–3401.

Hadfield, S.T., Hole, S.J.W., Howe, P.W.A., Stanley, P.D., 2001. Metabolite profiling by NMR for high-throughput mode of action identification of screen hits. Weeds 2, 551–556.

Hiltunen, Y., Heiniemi, E., Ala-Korpela, M., 1995. Lipoprotein-lipid quantification by neural-network analysis of ^{1}H NMR data from human blood plasma. J. Magn. Reson. B 106, 191–194.

Hole, S.J.W., Howe, P.W.A., Stanley, P.D., Hadfield, S.T., 2000. Pattern recognition analysis of endogenous cell metabolites for high throughput mode of action identification: removing the postscreening dilemma associated with whole-organism high throughput screening. J. Biomol. Screen. 5, 335–342.

Holmes, E., Foxall, P.J.D., Neild, G.H., Beddell, C., Sweatman, B.C., Rahr, E., Lindon, J.C., Spraul, M., Nicholson, J.K., 1994. Automatic data reduction and pattern recognition methods for analysis of ^{1}H nuclear magnetic resonance spectra of human urine from normal and pathological states. Anal. Biochem. 220, 284–296.

Jackson, M., Mantsch, H.H., 1996. Biomedical Infrared Spectroscopy. In: Mantsch, H.H., Chapman, D. (Eds.), Infrared Spectroscopy of Biomolecules. Wiley-Liss, New York, pp. 311–340.

Jackson, M., Mansfield, J.R., Dolenko, B., Somorjai, R.L., Mantsch, H.H., Watson, P.H., 1999. Prediction of breast tumor grade and steroid receptor status by multivariate analysis of Fourier transform infrared spectra. Cancer Detection and Prevention 23, 245–253.

Lee, H.-S., Chung, Y.H., Kim, C.Y., 1991. Specificities of serum alphafetoprotein in HBsAg+ and HBsAg- patients in the diagnosis of hepatocellular carcinoma. Hepatology 14, 68–72.

Lisboa, P. J. G, Branston, N. M., El-Deredy, W., Vellido, A., 1997. Assessment of statistical and neural networks methods in NMR spectral classification and metabolite selection. In: Proceedings of the IEEE/INNS International Joint Conference on Neural Networks, Houston, pp. 1385–1390.

Lisboa, P.J.G., Kirby, S.P.J., Vellido, A., Lee, Y.Y.B., El-Deredy, W., 1998. Assessment of statistical and neural networks methods in NMR spectral classification and metabolite selection. NMR in Biomedicine 11, 225–234.

Lutterbach, R., Stöckigt, J., 1995. Dynamics of the biosynthesis of methylursubin in plant cells employing in vivo ^{13}C NMR without labeling. Phytochemistry 40, 801–806.

Lutterbach, R., Stöckigt, J., 1994. In vivo investigation of plant-cell metabolism by means of natural-abundance C-13-NMR spectroscopy. Helvetica Chimica Acta 77, 2153–2161.

Mansfield, J.R., Sowa, M.G., Scarth, G.B., Somorjai, R.L., Mantsch, H.H., 1997. Analysis of spectroscopic imaging data by fuzzy C-means clustering. Anal. Chem. 69, 3370–3374.

Matsumoto, I., Kuhara, T., 1996. A new chemical diagnostic method for inborn errors of metabolism by mass spectrometry—rapid, practical, and simultaneous urinary metabolites analysis. Mass Spectrom. Rev. 15, 43–57.

Nicholson, J.K., Sadler, P.J., Bales, J.R., Juul, S.M., MacLeod, A.F., Sonken, P.H., 1984. Monitoring metabolic diseases by proton NMR of urine. Lancet 2, 751–752.

Nicholson, J.K., Wilson, I.D., 1989. High resolution proton magnetic resonance spectroscopy of biological fluid. Prog. NMR Spectr. 21, 449–501.

Nishijima, T., Fujiwara, K., 1997. Measurement of lactate levels in serum and bile using proton nuclear magnetic resonance in patients with hepatobiliary diseases: its utility in detection of malignancies. Jpn. J. Clin. Oncology 27, 13–17.

Ohsaka, A., Yoshikawa, K., Matsuhashi, T., 1979. Detection by proton nuclear magnetic resonance of elevated lactate concentration in serums from patients with malignant tumors. Jpn. J. Med. Sci. Biol. 32, 305–309.

Petroff, O.A.C., 1988. Biological ^{1}H NMR spectroscopy. Comp. Biochem. Physiol. 90B (2), 249–260.

Pope, J.M., Jonas, D., Walker, R.R., 1993. Applications of NMR microimaging to the study of grape berries. Protoplasma 173, 177–186.

Prabhu, V., Chatson, K.B., Abrams, G.D., King, J., 1996. ^{13}C Chemical shifts of 20 free amino acids and their use in detection by NMR of free amino acids in intact plants. J. Plant Physiol. 149, 246–250.

Rabenstein, D.L., Millis, K.K., Strauss, E.J., 1988. Proton NMR spectroscopy of human blood plasma and red cells. Anal. Chem. 60, 1380A–1391A.

Ratcliffe, R.G., Shachar-Hill, Y., 2001. Probing plant metabolism with NMR. Ann. Rev. Plant Physiol. Plant Mol. Biol. 52, 499–526.

Sauter, H., Lauer, M., Fritsch, H., 1991. Metabolic profiling of plants—a new diagnostic technique. In: Baker, D.R., Fenyes, J.G., Moberg, W.K. (Eds.), Synthesis and Chemistry of Agrochemicals II. ACS Symposium Series 443. American Chemical Society, Washington, DC, pp. 288–299.

Schmidt, R. R., 1997. HRAC classification of herbicides according to mode-of-action. In: Brighton Crop Protection Conference, Weeds, pp. 1133–1140.

Schneider, B., 1997. In vivo NMR spectroscopy of low-molecular compounds in plant cells. Planta 203, 1–8.

Shaw, R.A., Kotowich, S., Eysel, H.H., Jackson, M., Thomson, G.T.D., Mantsch, H.H., 1995. Arthritis diagnosis based upon the near-infrared spectrum of synovial fluid. Rheumatol. Int. 15, 159–165.

Somorjai, R.L., Dolenko, B., Nikulin, A.K., Pizzi, N., Scarth, G., Zhilkin, P., Halliday, W., Fewer, D., Hill, N., Ross, I., West, M., Smith, I.C.P., Donnelly, S.M., Kuesel, A.C., Brière, K.M., 1996. Classification of 1H NMR spectra of human brain neoplasms: the influence of preprocessing and computerized consensus diagnosis on classification accuracy. J. Magn. Reson. Imaging 6, 437–444.

Weckwerth, W., Fiehn, O., 2002. Can we discover novel pathways using metabolomic analysis? Curr. Opin. Biotechnol. 13, 156–160.

Wolfender, J.L., Hostettmann, K., 1996. LC-UV-MS: a powerful approach for the rapid screening of metabolites in crude plant extracts. In: Newton, R.P., Walton, T.J. (Eds.), Applications of Modern Mass Spectroscopy in Plant Science Research. Oxford Press, Oxford, pp. 216–221.

Mellerson, Kendra 7B08 (5)

From: Gakh, Yelena
Sent: Tuesday, August 05, 2003 2:33 PM
To: STIC-EIC1700
Subject: 09890973

Dear Kendra:

please order one more list:

4. TITLE: "Assessment of 1H NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of Arabidopsis thaliana"
AUTHOR(S): *Ward, Jane L.; Harris, Cassandra; Lewis, Jennie; Beale, Michael H.*
CORPORATE SOURCE: Department of Agricultural Sciences, IACR-Long Ashton Research Station, University of Bristol, Bristol, BS41 9AF, UK
SOURCE: **Phytochemistry (Elsevier) (2003), 62(6), 949-957**

Thank you,

Yelena

Biotech
QK861.P45

Yelena G. Gakh, Ph.D.

Patent Examiner
USPTO, cp3/7B-08
(703)306-5906

Available online at www.sciencedirect.com
SCIENCE DIRECT®
PHYTOCHEMISTRY
Phytochemistry 62 (2003) 949–957
www.elsevier.com/locate/phytochem

Assessment of ^{1}H NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of *Arabidopsis thaliana*

Jane L. Ward, Cassandra Harris, Jennie Lewis, Michael H. Beale*

IACR-Long Ashton Research Station, Department of Agricultural Sciences, University of Bristol, Long Ashton, Bristol BS41 9AF, UK

Received 26 August 2002; received in revised form 14 November 2002

Abstract

An approach to metabolite fingerprinting of crude plant extracts that utilizes ^{1}H nuclear magnetic resonance (NMR) spectroscopy and multivariate statistics has been tested. Using ecotypes of *Arabidopsis thaliana* as experimental material, a method has been developed for the rapid analysis of unfractionated polar plant extracts, enabling the creation of reproducible metabolite fingerprints. These fingerprints could be readily stored and compared by a variety of chemometric methods. Comparison by principal component analysis using SIMCA-P allowed the generation of residual NMR spectra of the compounds that contributed significantly to the differences between samples. From these plots, conclusions were drawn with respect to the identity and relative levels of metabolites differing between samples.

Keywords: *Arabidopsis thaliana*; NMR spectroscopy; Metabolomics; Multivariate analysis; Principal component analysis; Metabolite fingerprints

1. Introduction

Arabidopsis thaliana (*Arabidopsis*) is well known as a model system in plant research due to its relatively small genome, rapid life cycle, easy cultivation and high level of seed production. The completion of the genome sequence of *Arabidopsis* has provided the impetus for understanding the function of all the genes in this model plant (The *Arabidopsis* Genome Initiative, 2000; Wixon, 2001). Techniques such as proteomics and metabolomics may provide the necessary data to link gene sequence to function via the metabolic network (Hall et al., 2002; Fiehn, 2002) and thus high-throughput metabolomic analysis in *Arabidopsis* is an important goal in plant functional genomics (Trethewey, 2001). The "metabolome" has been defined, in a microbial context, as the total complement of metabolites in a cell (Tweeddale et al., 1998). For plants, examination of the metabolome is a more complex problem due to the larger number of potential metabolites and the presence of differentiated tissue, including specialist storage organs, with different metabolite complements. It is unlikely that a single analytical method will yield information about all the metabolites in a plant system. Differences due to volatility, polarity, solubility and chromatographic behaviour mean that multiple methods will need to be deployed to analyse different subsets of metabolites. In this context coupled gas chromatography–mass spectrometry (GC–MS) has already been successfully applied to plant metabolite profiling (Roessner et al., 2001), including *Arabidopsis* (Fiehn et al., 2000), where 326 distinct compounds from leaf extracts were quantified. Another potentially powerful tool for plant metabolite analysis is high-resolution nuclear magnetic resonance spectroscopy (NMR), in particular ^{1}H NMR. This technology has been utilized extensively to profile metabolites in clinical samples (e.g. Nicholson and Wilson, 1989; Holmes et al., 2000; Beckwith-Hall et al., 2002) and has been applied to complex mixtures of compounds exuded from cereal roots (Fan et al., 2001). Unlike GC–MS, which detects only those compounds that can be volatilized (usually achieved by derivatization), ^{1}H NMR can simultaneously detect all proton-bearing compounds in a sample. This covers most of the "organic" compounds such as carbohydrates, amino acids, organic and fatty acids, amines, esters, ethers and

* Corresponding author. Tel.: +44-1275-549289; fax: +44-1275-394281.
E-mail address: mike.beale@bbsrc.ac.uk (M.H. Beale).

0031-9422/03/$ - see front matter
doi:10.1016/S0031-9422(02)00705-7

lipids, which are present in plant tissues. Thus, NMR spectra of unpurified solvent extracts of plants has the potential to provide a relatively unbiased fingerprint, containing overlapping signals of the majority of the metabolites present in the solution.

In this paper we report the development of one-dimensional ^{1}H NMR spectroscopy methods, coupled with multivariate statistical analysis (Antti et al., 2002), for the analysis of crude extracts of *Arabidopsis*. A small set of ecotypes was used as suitable experimental material to develop the method, as a previous GC–MS study had shown that the metabolite profiles of two ecotypes, (Col-2 and C24) showed significant differences (Fiehn et al., 2000).

2. Results and discussion

2.1. Extraction and analysis of plants

The *Arabidopsis* ecotypes employed (Table 1) were grown under identical long-day controlled environment conditions in trays containing 24 individual plants. Plants were harvested at growth stage 6.1–6.5 (just bolted, first flower present) as described by Boyes et al. (2001). In order to smooth out plant to plant variability, all aerial plant material from each tray was combined. Enzyme activity was stopped by immediately immersing the harvested material in liquid nitrogen, before freeze-drying. The extraction method developed is relatively simple, requiring a suspension of weighed aliquots of the powdered, freeze-dried plant material in deuterated NMR solvent (80:20 D_2O:CD_3OD), a short period of moderate heating, followed by micro-centrifugation. An aliquot of the supernatant was then analysed directly by ^{1}H NMR. An obvious advantage of this method is the use of deuterated solvents for tissue extraction. This eliminated the need for evaporation and re-dissolution of extracts, which has the associated potential problem of loss of material. Three sample replicates were taken in each case to assess the robustness of the sample preparation method.

2.2. Features of ^{1}H NMR spectra of polar extracts of Arabidopsis

In general the NMR spectra obtained showed a dominance of signals in the carbohydrate region of the spectrum. In addition to these signals, well-defined signals in both the aromatic and aliphatic regions of the spectra were present (Fig. 1). The sharp singlet at δ 6.5 was identified as fumaric acid. Similarly, other signals in relatively clear areas of the trace could be assigned to

Table 1
Arabidopsis ecotypes used in assessment of multivariate analysis by ^{1}H NMR spectroscopy

Name of ecotype	Code	NASC code	Country of origin
Columbia	COL-0	N1092	USA
Landsberg	LER-1	N1642	Germany
Dijon	Di-0	CS1106	France
Estland	Est-0	N1148	Russia
Nossen	No-0	N3081	Germany
Wassilewskija	WS-0	N1602	Russia
Wassilewskija	WS-2	N1601	Russia
C24	C24	N906	USA
Rschew	Rld-2	N1641	Russia

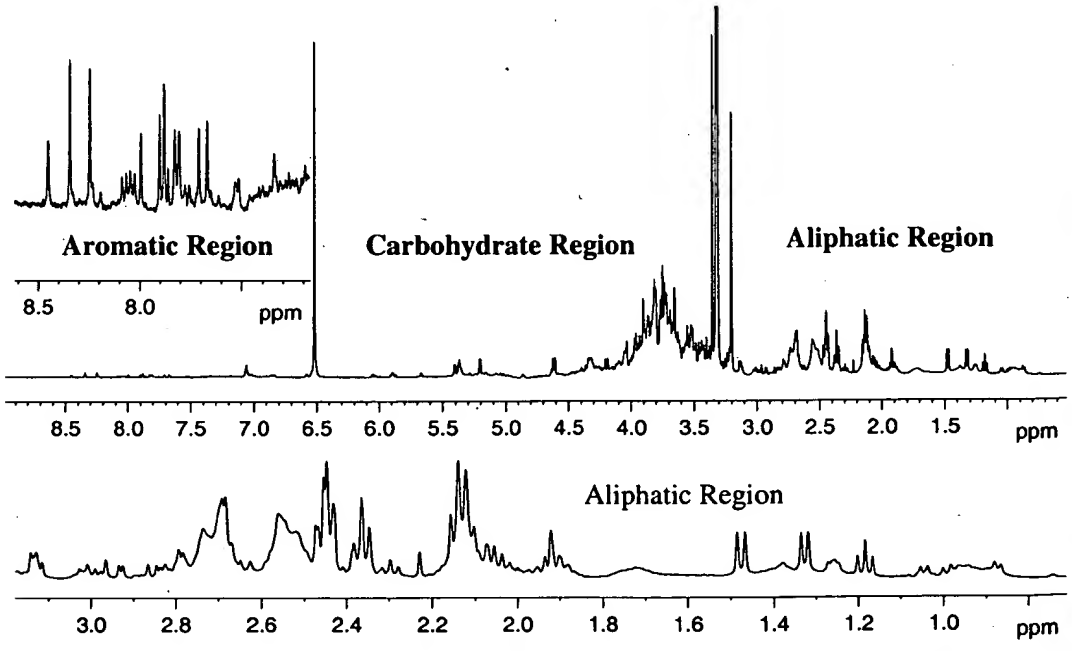


Fig. 1. ^{1}H NMR spectrum of a typical *Arabidopsis* (Landsberg) polar extract in D_2O:CD_3OD.

particular amino acids (e.g. alanine doublet at δ 1.47 and threonine doublet at δ 1.32), and particular carbohydrates (e.g. α- and β-glucose anomeric hydrogens at δ 5.20 and 4.60). From visual analysis of spectra from the nine ecotypes, clear differences were evident. For example, by comparison to the other eight ecotypes in the set, WS-0 had significantly increased intensity in many of the carbohydrate signals (Fig. 2). In addition, the ecotype Dijon possessed completely new signals in the region δ 6.0–4.90 (Fig. 2). Other differences included a variation in intensities of the same signals in different ecotypes (e.g. the fumaric acid signal).

2.3. Standardization and processing of the data for export

For electronic comparison of the data sets by multivariate methods it was important to ensure that there is as little experimental variation as possible in the sample set. The spectra were all Fourier-transformed, in automation, using the same processing parameters, an exponential window and a line-broadening factor of 0.5 Hz. Each data set was automatically scaled to trimethylsilylpropionate-d_4 internal standard, phased and baseline corrected. After importation into AMIX (Analysis of MIXtures software, Bruker, Germany) the negative peaks were removed and a compressed form of the data was stored in a spectral database for future reference. Data from standards were collected and processed in an identical fashion and stored in the AMIX spectral database. Before analysis by multivariate methods, data sets, selected from the database, were reduced in complexity by using the "bucketing" function to generate a set number of integrated regions or "bins" of the data set. This table of 'binned' data from those spectra selected could then be exported as a spreadsheet suitable for importation into statistical analysis software, such as SIMCA-P (Umetrics). The ability to batch process datasets from any number of samples, held in the database, as described represents a further benefit of using ^{1}H NMR to collect metabolite fingerprints. Currently some other methodologies for large-scale metabolite fingerprinting, for example GC–MS, the ability to database aligned and normalised data

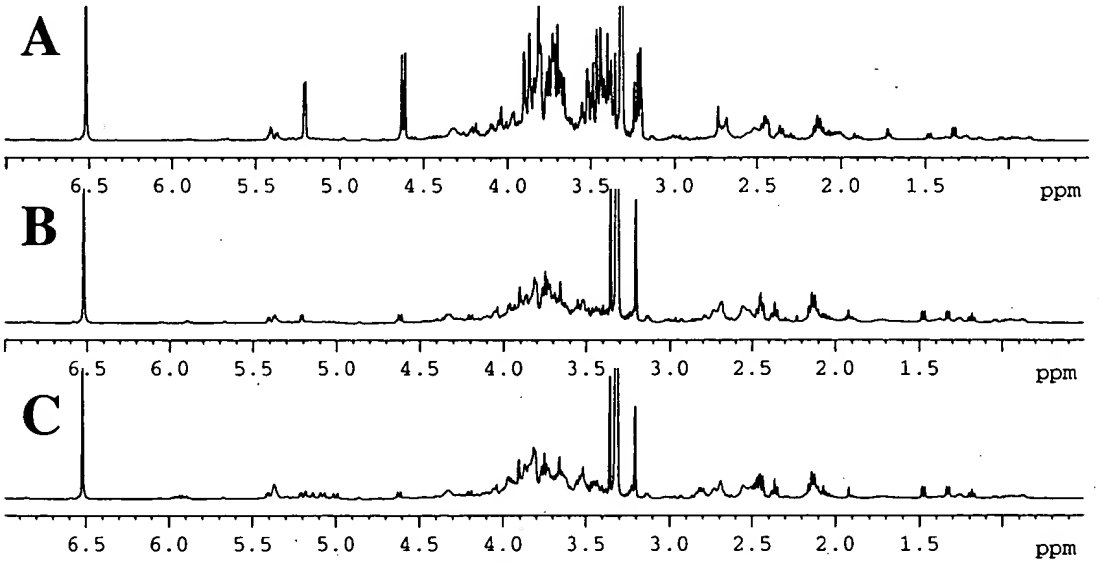


Fig. 2. ^{1}H NMR spectra of three *Arabidopsis* ecotypes. A: WS-0, B: Landsberg, C: Dijon.

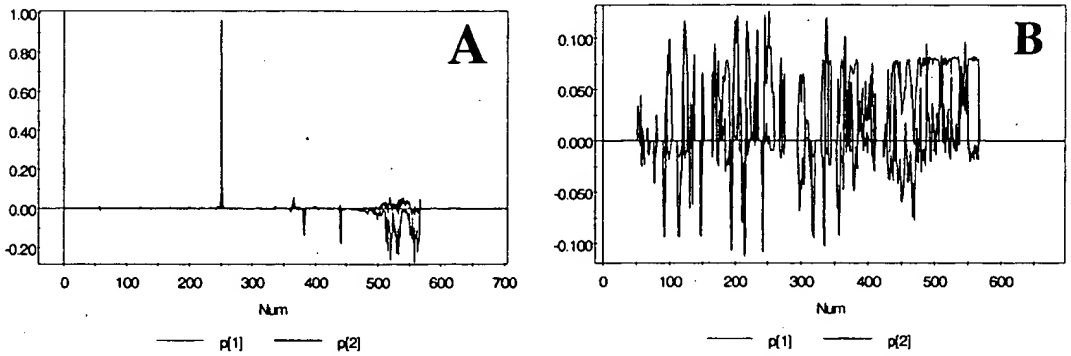


Fig. 3. Comparison of typical loadings plots generated using A: the covariance matrix and B: the correlation matrix.

sets *in automation* is not easily achieved within current spectrometer operating software. The inherent problems of retention time drift, and column and source variability, mean that peak alignment and data export methods from GC–MS require operator quality control to ensure accurate peak alignment to prepare each data set for alignment, storage and multivariate analysis.

2.4. Principal component analysis (PCA) of *Arabidopsis* ecotype data sets

PCA is a data visualization method that is useful for observing groupings within multivariate data. Data is represented in n dimensional space, where n is the number of variables, and is reduced into a few principal components, which are descriptive dimensions that describe the maximum variation within the data. The principal components can be displayed in a graphical fashion as a "scores" plot. This plot is useful for observing any groupings in the data set and in addition will highlight outliers that may be due to errors in sample preparation or instrumentation parameters. PCA models are constructed using all the samples in the study. Coefficients by which the original variables must be multiplied to obtain the PC are called "loadings." The numerical value of a loading of a given variable on a PC shows how much the variable has in common with that component (Massart et al., 1988). Thus for NMR data, "loading plots" can be used to detect the spectral areas responsible for the separation in the data.

The data for PCA can be scaled in different ways. If the data is mean-centred with no scaling then a covariance matrix is produced, but if the data mean-centred and the columns of the data matrix scaled to unit variance, a correlation matrix is produced. An advantage of the covariance matrix is that the loadings retain the scale of the original data. In the case of the data reported here, the loadings plots, when viewed as line plots, resemble NMR spectra and can be interpreted as such (Fig. 3). In contrast, the correlation matrix produces loadings plots which are unfamiliar in appearance (Fig. 3). For the purposes of this work, a covariance matrix was used to allow for a more useful interpretation of the loadings plots. Contribution plots allow further interpretation of the differences observed in the scores diagram, and depict the changes in variables (e.g. chemical shift) between two observations (samples) or between a selected observation and the average. When plotted as line diagrams these also resemble NMR spectra and in that sense depict spectra of compounds responsible for the differences between chosen samples.

For the data set obtained from replicate analysis of the ecotypes, a nine-component model explained 99% of the variance, with the first two components explaining

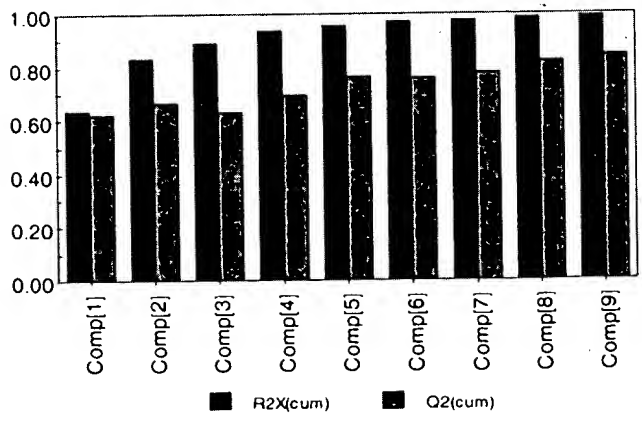


Fig. 4. Model overview illustrating the number of components and explained variances used in PCA analysis of *Arabidopsis* ecotypes.

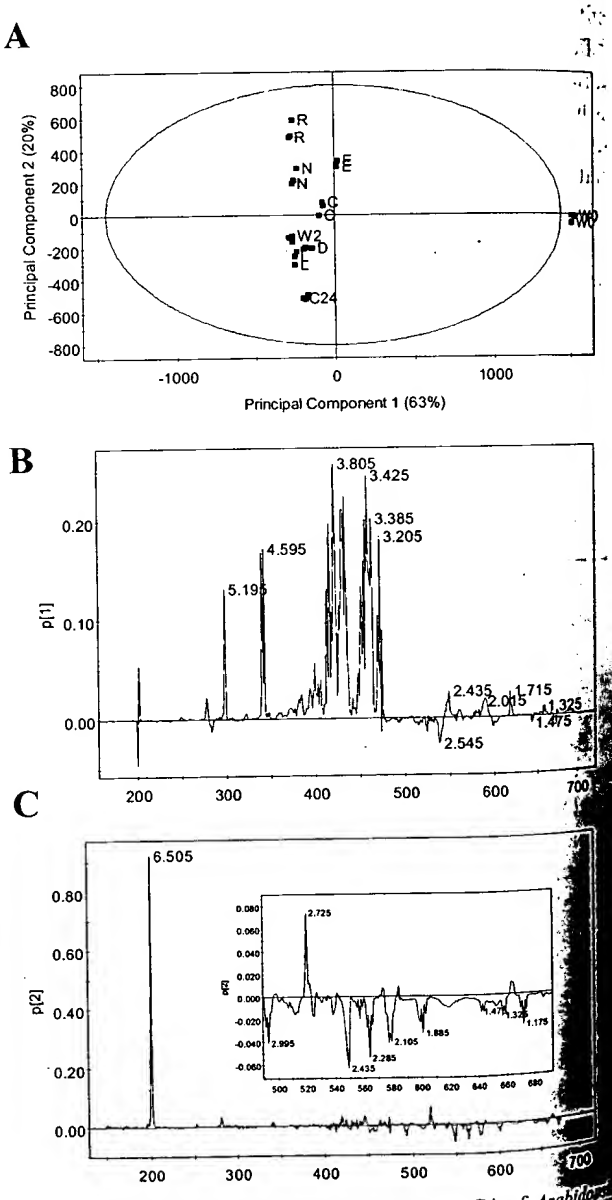


Fig. 5. Scores and loadings plots generated from PCA of *Arabidopsis* ecotypes. A: scores plot of PC1 vs. PC2. B: loadings plot of PC1. C: loadings plot of PC2. W0 = WS-0, W2 = WS-2, E = Estland, D = Dijon, R = Rschew, N = Nossen, C = Columbia-0, L = Landsberg, C24 = C24.

of the variability (Fig. 4). Examination of the scores and loadings plots for PC1 vs. PC2 (Fig. 5) showed good experimental replication since tight clustering of replicate samples could be seen with several of them clustering on top of each other. Examination of the scores plot (Fig. 5A) demonstrated that WS-0 was separate from the rest of the group. Examination of the loadings plot of PC1 (Fig. 5B) showed that the first component explained the variance in carbohydrate levels since high loadings values were observed for peaks in the carbohydrate region of the NMR spectrum. In addition, the loadings plot of PC1 illustrated some small positive and some small negative regions of the spectrum between δ 2.5 and δ 1.25 (Fig. 5C). This region contained many peaks attributable to amino acids and this information may give clues as to the variance of amino acids between ecotypes. It is evident that WS-0 is separated mainly by virtue of its increase in carbohydrate relative to the rest of the group. Examination of the scores plot (Fig. 5A) also indicated that the rest of the set of ecotypes had fairly similar levels of carbohydrate. In order to correctly determine the nature of this increased carbohydrate we examined, in AMIX, the spectrum of WS-0 against a library of spectra of carbohydrates run in the same solvent under the same conditions. As can be seen in the contribution plot, Fig. 6A, the increased peaks were due to glucose (approx. 1:1 anomeric mixture). Thus it would appear that WS-0 has elevated levels of glucose relative to all of the other ecotypes examined here. This observation was confirmed by the quantitative GC–MS analysis of methoxyamine-terimethylsilylated samples relative to added ribitol internal standard (Roessner et al., 2001). The results indicated that glucose levels in WS-0 were four times higher than in the other ecotypes, while other simple carbohydrates, such as fructose, mannose and galactose, which are of lower abundance than glucose,

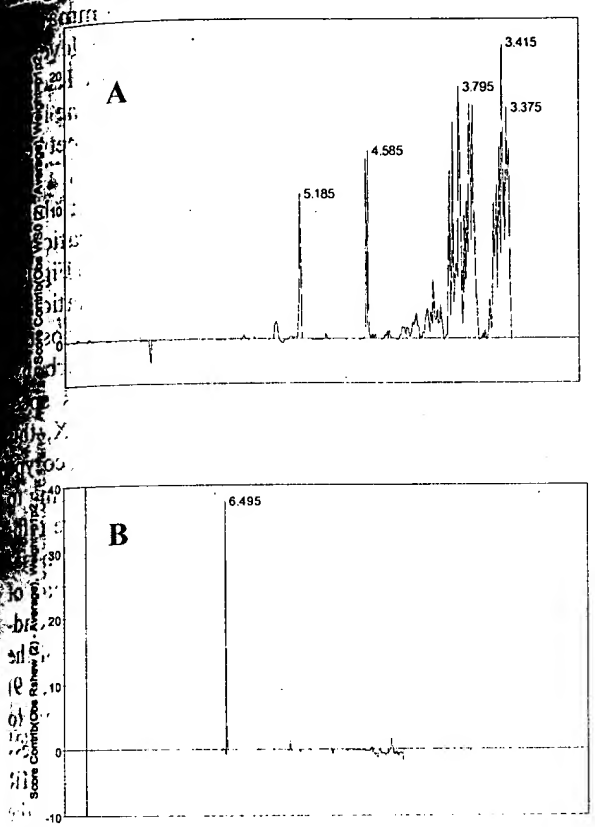


Fig. 6. Contribution plots of A: WS-0 *minus* average and B: C24 *minus* average, generated from PC1 vs. PC2 scores plot from PCA of *Arabidopsis* ecotypes.

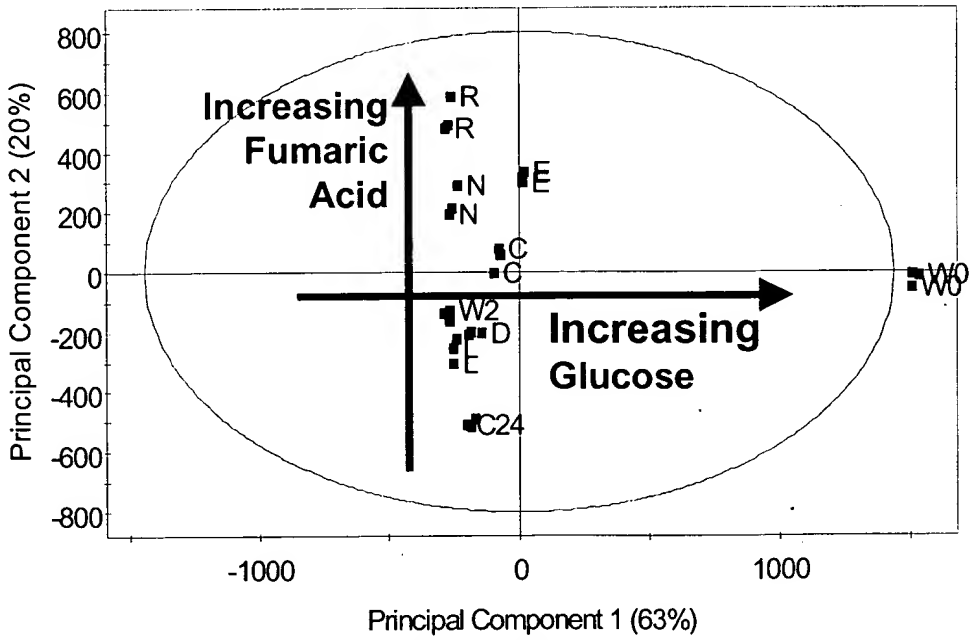


Fig. 7. Analysis and summary of the scores plot of PC1 vs. PC2, after examination of the associated loadings plots, indicating the metabolites responsible for the greatest variance.

were also elevated to 2–3 times those of the other ecotypes. The origin of the highly elevated monosaccharides in WS-0 is unclear at present. Ecologically, the origin of WS-0 resembles WS-2. It is possible that the high sugar levels are a result of increased polysaccharide hydrolytic activity in this ecotype. The possibility that this kind of enzyme activity may be manifested during sample processing was investigated by repeating the extraction procedure several times and by re-running the NMR spectra after storage of the samples. No indication of such post-harvest degradation was found, but further experiments are necessary to fully investigate this. The loadings plot (Fig. 5C) and the contribution plot (Fig. 6B) of the second principal component PC2 was relatively simple with a large peak at δ 6.5. This peak has been identified as fumaric acid by the comparison of a set of organic acid standards run in the same solvent system and confirmed by addition of fumaric acid to an *Arabidopsis* NMR sample. The scores plot of PC1 vs. PC2 can now be summarized according to Fig. 7, which shows that the level of fumaric acid in the set of ecotypes varies with Rschew having the highest amount and C24 possessing the least. Columbia fumarate levels are intermediate between these.

PC3 and PC4 accounted for the next 9% of variability within the sample set and demonstrated a separation of Estland and Dijon from the rest of the group (Fig. 8). Estland was separated by virtue of PC3. Examination of the loadings plot for PC3 (Fig. 8B) shows positive loadings for some (non-glucose) signals in the carbohydrate region. Examination of the original NMR spectrum and by comparison of standards in AMIX, this was identified as the disaccharide maltose. The ecotype Dijon separated from the rest of the group according to PC4. The loadings plot for PC4 (Fig. 8C) is more difficult to interpret since there are both positive and negative loadings. Since Dijon was found in the lower half of the scores plot we can infer that it is the negative loadings that are associated with Dijon. Examination of the contribution plots for Estland and Dijon (Fig. 9) revealed approximate NMR spectra corresponding to increased metabolites that these two ecotypes possess over the rest of the group. New signals in the ^{1}H NMR spectrum for Dijon were observed, and through the analysis of the contribution plot for "Dijon minus average", clues to the identity of this compound(s) are revealed. The compound appears to be olefinic or contain an unsaturated heterocyclic ring. So far the comparison of Dijon with metabolite standards and further investigation by two-dimensional NMR, and GC–MS, has yet to reveal the identity of this metabolite. On the other hand the contribution plot of Estland confirmed the presence of elevated levels of maltose and several amino acids, including lysine (δ 1.92).

Examination of the higher PCs (PC5–PC9) highlighted further differences in the sample set (data not shown). For example, PC5 vs. PC6 separated WS-2, and analysis of PC7 vs. PC8 separated the Landsberg ecotype. Differences in amino acids such as valine, isoleucine and threonine could be detected by examination of these PCs and further inspection of the original NMR spectra confirmed these minor differences.

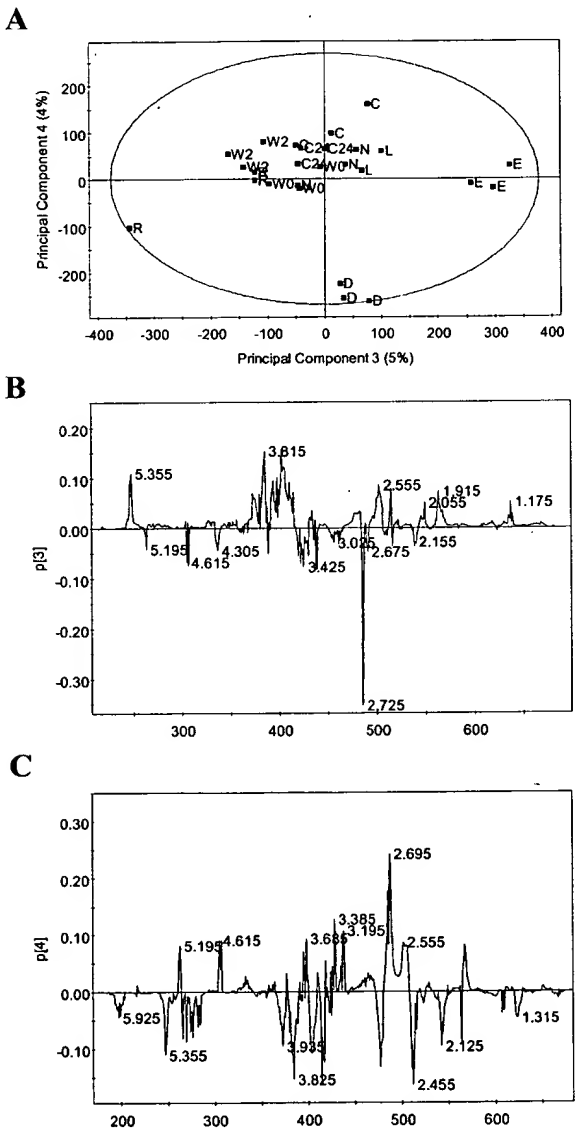


Fig. 8. Scores and loadings plots generated from PCA of *Arabidopsis* ecotypes. A: scores plot of PC3 vs. PC4. B: loadings plot of PC3. C: loadings plot of PC4. W0 = WS-0, W2 = WS-2, E = Estland, D = Dijon, R = Rschew, N = Nossen, C = Columbia-0, L = Landsberg, C24 = C24.

2.5. Reproducibility in the method

It can be seen from the scores plot that the experimental variability is acceptable since tight distinct clusters form corresponding to each ecotype. The whole procedure, from extraction to data analysis, was repeated on the same set of freeze-dried plant samples

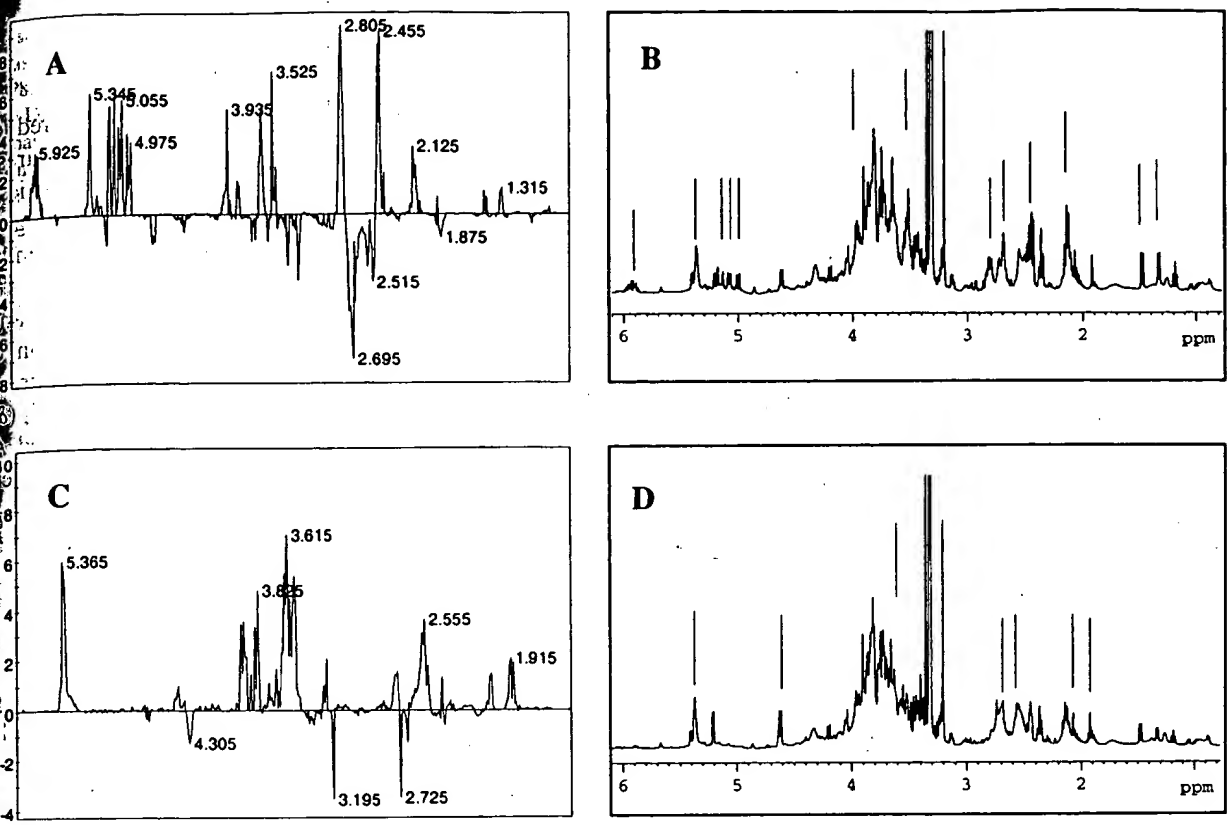


Fig. 9. Contribution plots generated from PCA of *Arabidopsis* ecotypes. A: contribution plot of Dijon *minus* average. B: ^{1}H NMR spectrum of Dijon. C: contribution plot of Estland *minus* average. D: ^{1}H NMR spectrum of Estland.

order to determine the reproducibility of the method as a whole. When the data were modelled using PCA in the same way, clustering of each ecotype was observed in a similar fashion to that seen previously. There was no separation of the individual clusters, again indicating the method was reproducible. Relative standard deviations were calculated for each observation of each ecotype, in the ^{1}H NMR spectrum. The mean of these deviations was 12±4%.

The earlier experiments utilized aliquots of combined material from trays of plants, all grown at the same time in controlled environment. In these experiments plant to plant biological variability was not assessed. However, examination of extracts from single Landsberg plants by the method above (data not shown) indicated that plant-to-plant variability was quite large. The mean relative standard deviation was 52±7%. In a another experiment aliquots of combined freeze-dried tissue from replicate trays were analysed. In this case the mean relative standard deviation was calculated to be 28±3%. These results indicate that pooling of plants grown together can reduce differences in data sets due to biological variability. However variability due to effects such as position of trays in the growth chamber is still significant.

3. Conclusions

^{1}H NMR spectroscopy has proved to be a valuable tool for unbiased metabolite fingerprinting of *Arabidopsis*. Principal component analysis highlighted genuine differences between ecotypes with loadings plots giving clues as to the nature of these differences. Comparison of the spectra of highlighted ecotypes with a library of NMR spectra of standards run under identical conditions, in AMIX, allowed us to identify compounds responsible for differences between spectra of different ecotypes. Differences could be detected in both the carbohydrate region and the aliphatic region, with sugars, organic acids and amino acids contributing to the differences in the sample set. The work has demonstrated how ^{1}H NMR analysis may be used in the future as a first pass screen to rapidly determine and characterize differences in molecular composition of plant samples. The technique serves as a rapid fingerprinting method that compares favourably with FT-IR (Goodacre and Anklam, 2001) with respect to reproducibility and extent of metabolome coverage. NMR, however, has the advantage over FT-IR in that the identities of many of the major metabolites can be deduced from the spectra. Coupled-MS techniques have advantages both

in terms of numbers of metabolites that can be quantified and the dynamic range of the concentrations that can be measured, but suffer from the disadvantage that chromatography selects subsets of the total metabolites. An integrated approach where differences are thrown up by NMR screening and then further investigated and accurately quantified by more targeted (chromatography-linked) methods such as GC–MS with appropriate internal standards seems to be a reasonable way forward to initiate high-throughput screens of plants. In this respect we foresee many uses of the NMR technique described, from large-scale analysis of natural variation, through mutant collections to transgenic plants.

4. Experimental

4.1. Plant material

Arabidopsis thaliana seeds were obtained from Nottingham *Arabidopsis* Seed Centre (NASC) and were germinated on agar containing Gamberg's B-5 basal medium containing 3% sucrose at 22 °C in continuous light. Plants were transferred to soil at the 2–4 leaf stage and grown in a controlled environment under long day (16 h) conditions, at a temperature of 23 °C and 75% humidity during the day and 18 °C and 80% humidity at night. Plants were harvested at growth stage 6.1–6.5 (Boyes et al., 2001) and immediately plunged into liquid nitrogen before freeze drying and grinding to a fine powder in a pestle and mortar. Samples were then stored until required at −80 °C.

4.2. Extraction and ^1H NMR spectroscopy

Freeze-dried plant material (15 mg) was weighed into an autoclaved 2 ml Eppendorf tube. D_2O:CD_3OD (1 ml, 80:20) containing 0.05% w/v TSP-d_4 (sodium salt of trimethylsilylpropionic acid) was added to each sample. The contents of the tube were mixed thoroughly and then heated at 50 °C in a water bath for 10 min. After cooling, the samples were spun down in a micro-centrifuge for 5 min. Of the supernatant 750 μl were added to a 5 mm NMR tube. All spectra were acquired under automation at a temperature of 300 K on a Bruker Avance spectrometer operating at 399.752 MHz ^1H observation frequency using the multinuclear broadband BBO 5 mm probe, and a water suppression pulse sequence with a relaxation delay of 5 s. Each spectrum consisted of 2048 scans of 32 k data points with a spectral width of 4845 Hz. The spectra were automatically Fourier transformed using an exponential window with a line broadening value of 0.5 Hz, phased and baseline corrected within the automation programme. ^1H NMR chemical shifts in the spectra were referenced to TSP-d_4 at δ 0.00.

4.3. Data reduction of the NMR spectra and multivariate analysis

The ^1H NMR spectra were automatically reduced to ASCII files using AMIX (Analysis of MIXtures software v.3.0, Bruker Biospin). Spectral intensities were scaled to TSP-d_4 and reduced to integrated regions or "buckets" of equal width (0.01 ppm) corresponding to the region of δ 9.0 to δ −0.5. The regions between δ 4.90 and δ 4.76 were removed prior to statistical analyses thus eliminating any variability in suppression of the water sample. The residual proton signals corresponding to methanol-d_4 (δ 3.365–3.285) and TSP-d_4 (δ 0.00) were also removed at this stage. The generated ASCII file was imported into Microsoft EXCEL for the addition of labels and then imported into SIMCA-P 9.0 (Umetrics, Umea, Sweden) for PCA analysis.

Acknowledgements

IACR-Long Ashton Research Station receives grant-aided support from the Biotechnology and Biological Sciences Research Council of the United Kingdom. We thank the BBSRC-GARNet project (http://www.york.ac.uk/res/garnet/garnet.html) for support (C.H. and J.L.).

References

Antti, H., Bollard, M.E., Ebbels, T., Keun, H., Lindon, J.C., Nicholson, J.K., Holmes, E., 2002. Batch statistical processing of ^1H NMR-derived urinary spectral data. J. Chemometrics 16, 461–468.

Beckwith-Hall, B.M., Holmes, E., Lindon, J.C., Gounarides, J., Vickers, A., Shapiro, M., Nicholoson, J.K., 2002. NMR-based metabonomic studies on the biochemical effects of commonly used drug carrier vehicles in the rat. Chem. Res. Toxicol. 15, 1136–1141.

Boyes, D.C., Zayed, A.M., Ascenzi, R., McCaskill, A.J., Hoffman, N.E., Davis, K.R., Görlach, J., 2001. Growth stage-based phenotypic analysis of *Arabidopsis*: a model for high throughput functional genomics in plants. Plant Cell 13, 1499–1510.

Fan, T.W.-M., Lane, A., Shenker, M., Bartley, J.P., Crowley, D., Higashi, R.M., 2001. Comprehensive chemical profiling of gramineous plant root exudates using high-resolution NMR and MS. Phytochemistry 57, 209–221.

Fiehn, O., Altmann, T., Trethewey, R.N., Willmitzer, L., 2000. Metabolite profiling for plant functional genomics. Nature Biotech. 18, 1157–1161.

Fiehn, O., 2002. Metabolomics: the link between genotypes and phenotypes. Plant Mol. Biol. 48, 155–171.

Goodacre, R., Anklam, E., 2001. Fourier transform infrared spectroscopy and chemometrics as a tool for the rapid detection of other vegetable fats mixed in cocoa butter. J. Am. Oil Chem. Soc. 78, 999–1000.

Hall, R., Beale, M.H., Fiehn, O., Hardy, N., Sumner, L., Bino, R., 2002. Plant metabolomics: the missing link in functional genomics strategies. Plant Cell 14, 1437–1440.

Holmes, E., Nicholl, A.W., Lindon, J.C., Connor, S.C., Connelly, J.C., Haseldon, J.N., Damment, S.J.P., Spraul, M., Neidig, P.

Nicholson, J.K., 2000. Chemometric methods for toxicity classification based on NMR spectra of biofluids. Chem. Res. Toxicol. 13, 471–478.

Massart, D.L., Vandeginste, B.G.M., Deming, S.N., Michotte, Y., Kauffman, L., 1988. Chemometrics: A Textbook. Elsevier, New York.

Nicholson, J.K., Wilson, I.D., 1989. High-resolution proton magnetic resonance spectroscopy of biological fluids. Prog. Nucl. Magn. Reson. Spectrosc. 21, 449–501.

Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L., Fernie, A., 2001. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. Plant Cell 13, 11–29.

The *Arabidopsis* Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408, 796–815.

Trethewey, R.N., 2001. Gene discovery via metabolite profiling. Curr. Opin. Biotechnol. 12, 135–138.

Tweeddale, H., Notley-McRobb, L., Ferenci, T., 1998. Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool ("metabolome") analysis. J. Bacteriol. 180, 5109–5116.

Wixon, J., 2001. Featured organism: *Arabidopsis thaliana*. Comp. Funct. Genom. 2, 91–98.

Mellerson, Kendra

From: Gakh, Yelena
Sent: Tuesday, August 05, 2003 2:33 PM
To: STIC-EIC1700
Subject: 09890973

Dear Kendra:

please order one more list:

5. TITLE: "Multi-component metabolic classification of commercial feverfew preparations via high-field 1H-NMR spectroscopy and chemometrics"
AUTHOR(S): *Bailey, Nigel J. C.; Sampson, Julia; Hylands, Peter J.; Nicholson, Jeremy K.; Holmes, Elaine*
CORPORATE SOURCE: Biological Chemistry, Biomedical Sciences Division,
Imperial College of Science, Technology and Medicine, University of London, London, SW7 2AZ, UK
SOURCE: **Planta Medica (2002), 68(8), 734-738**

Thank you,

Yelena

Yelena G. Gakh, Ph.D.

Patent Examiner
USPTO, cp3/7B-08
(703)306-5906

Biotech
RS164.P7
or Adm

Article title	Multi-component metabolic classification of commercial feverfew preparations via high-field ^1H-NMR spectroscopy and chemometrics
Article identifier	0032094302002168
Authors	Bailey_N_J_C Sampson_J Hylands_P_J Nicholson_J_K Holmes_E
Journal title	Planta Medica
ISSN	0032-0943
Publisher	Thieme
Year of publication	2002
Volume	68
Issue	8
Supplement	0
Page range	734-738
Number of pages	5
User name	Adonis
Cost centre	
PCC	$21.00
Date and time	Thursday, August 07, 2003 2:13:35 AM

Multi-Component Metabolic Classification of Commercial Feverfew Preparations *via* High-Field ^{1}H-NMR Spectroscopy and Chemometrics

Nigel J. C. Bailey[1]
Julia Sampson[2,4]
Peter J. Hylands[3]
Jeremy K. Nicholson[1]
Elaine Holmes[1]

Original Paper

Abstract

There is increasing interest in evaluating the clinical efficacy of herbal medicines. However, there are significant analytical problems associated with quality control and the measurement of the overall composition of such complex, multi-component mixtures as normally required in the pharmaceutical industry. Here we describe a novel NMR spectroscopic and pattern recognition analytical approach to investigate composition and variability of a commonly used herbal medicine. 600 MHz ^{1}H-NMR spectroscopy and principal components analysis (PCA) was used to discriminate between batches of 14 commercially available feverfew samples based on multi-component metabolite profiles. Two of the batches were significantly different from the other twelve. The twelve remaining classes could be classified into discrete groups by PCA on the basis of minor differences in overall chemical composition. NMR based pattern recognition (PR) analysis of extracts proved to be superior to PR analysis of HPLC traces of the same mixtures.This work indicates the potential value of NMR combined with PCA for the characterisation of complex natural product mixtures, and the discrimination of samples containing allegedly identical ingredients.

Key words
Feverfew · NMR spectroscopy · pattern recognition · principal components analysis · quality control · sample classification *Tanacetum parthenium* · Compositae

Abbreviations
PCA: principal components analysis
PC: principal component
PR: pattern recognition
TSP: 3-(trimethylsilyl)-propionic-2,2,3,3-d_4 acid, sodium salt

Introduction

There is an increasing interest in the efficacy of many herbal medicines that have been used as natural remedies to treat a variety of ailments for centuries. This growing interest brings with it the need to develop analytical techniques capable of rapid and efficient analysis of these biologically complex 'single chemical entities'. The sheer complexity of these samples means that analysis for overall composition and quality control determinations are beyond the scope of more traditional pharmaceutical methods of analysis.

Here we report the application of high-field ^{1}H-NMR spectroscopy and multivariate data analysis to investigate the composition of feverfew sample batches. Feverfew [*Tanacetum parthenium* (L.) Schultz Bip (Compositae)] is a member of the daisy family that has been used as a natural headache remedy for centuries. Controlled clinical studies have shown that fever-

Affiliation
[1] Biological Chemistry, Biomedical Sciences Division, Imperial College of Science, Technology and Medicine, University of London, South Kensington, London, United Kingdom
[2] Department of Pharmacy, Franklin-Wilkins Building, King's College London, Waterloo, London, United Kingdom
[3] Oxford Natural Products plc, Cornbury Park, Charlbury, Oxfordshire, United Kingdom
[4] Now at Oxford Natural Products plc

Correspondence
Dr. Nigel J.C. Bailey · Biological Chemistry · Biomedical Sciences Division · Imperial College of Science, Technology and Medicine · University of London · Sir Alexander Fleming Building · Exhibition Road · South Kensington · London, SW7 2AZ · United Kingdom · E-Mail: nigel.bailey@ic.ac.uk · Fax: +44 020 7594 3226

Received October 30, 2001 · **Accepted** February 3, 2002

Bibliography
Planta Med 2002; 68: 734–738 · © Georg Thieme Verlag Stuttgart · New York · ISSN 0032-0943

few significantly reduces the frequency of migraine headaches [1].

The activity of feverfew is considered to be due to the presence of sesquiterpene lactones, principally the germacranolide, parthenolide. Parthenolide has been shown to inhibit the release of serotonin *in vitro*, which may be relevant to its effectiveness on migraine [2]. It has also been shown that the effectiveness of feverfew to prevent migraine is well correlated to levels of parthenolide within a sample [3].

Because of the assumed importance of parthenolide content on the efficacy of a particular feverfew sample, several methods have been developed for the analysis of parthenolide in feverfew samples utilising either HPLC or NMR spectroscopic methodologies [2], [4]. These methods have allowed the determination of different levels of parthenolide in feverfew samples of different sources. Such approaches, however, ignore the presence of other active sesquiterpene lactones [5], which may contribute to the observed clinical efficacy of the plant either synergistically (positive effects) or sysergistically (negative effects) [6].

In order to observe the overall gross differences between samples of different sources, it is necessary to apply some kind of 'chemometric' analysis to data representing the chemical composition of the samples as a whole. Chemometrics is a technology for exploring and modelling complex and often unknown relations in multivariate data. By applying such techniques (i.e., pattern recognition, PR) to the data, and using visual analysis, it is possible to elucidate hidden relations within the data [7].

Application of chemometric analyses to HPLC data of plant extracts has enabled the various feverfew and related species to be differentiated in terms of their gross chemical composition of substances with strong UV chromophores [8], [9]. HPLC derived data are, however, selective and depend on the choice of mobile and stationary phases as well as the wavelength employed for detection. Feverfew leaves are known to contain many classes of compound with potential biological activity, e.g., flavonoids [10], [11]. This means that standardisation methodologies relying on the attempted control of just one class of secondary metabolite are unlikely to represent true classification in terms of the "global biochemical makeup" in a reliable manner and in a way that correlates with the total clinical effect.

NMR-based pattern recognition (NMR-PR) analysis of complex biomixtures has been widely applied in the field of metabonomics (the study of changes in the metabolic profile of an organism as a whole in response to external influence) for characterising and predicting altered metabolic profiles from toxicological screening [12], [13]. However, applications of NMR-PR are extensive and have included discrimination between apple varieties [14] and grape cultivars [15].

One approach to pattern recognition commonly employed is principal components analysis (PCA). PCA condenses the multivariate data (i.e., NMR spectra) into a reduced number of orthogonal components that describe the greatest amount of variance in the data. This allows visual representation of the similarities (or differences) between samples within the dataset [16], [17]. This work reports the application of a combination of ^{1}H-NMR spectroscopy and PCA to distinguish between a number of commercially available feverfew samples, with a view to developing quality control procedures to guarantee the reproducibility of feverfew samples. The work demonstrates that NMR spectroscopy is the ideal analytical tool for such discrimination, as the non-selective nature of the technique means that the resultant spectrum is a true representation of the sample as a whole, and so subtle differences within the whole sample may be observed rapidly. Further, NMR based pattern recognition (PR) analysis of extracts was shown to be superior to PR analysis of HPLC traces of the same mixtures.

Materials and Methods

Sample preparation

Samples were obtained from different brands of feverfew available commercially. A voucher specimen (CH156 I1) was deposited in the herbarium of the Department of Pharmacy, King's College London UK.

Depending on availability, samples were run in duplicate or triplicate. Samples of 4.5 g of tablets were weighed accurately and ground for 15 minutes in a 'Moulinex' grinder. 100 ml of double distilled water (room temperature) were added to the powder in a conical flask and shaken at 150 rpm for 4 hours at room temperature. Extracts were filtered through a Whatman No.1 filter, the filtrate collected and freeze dried. Samples were then lyophilised and 10 mg of each sample reconstituted in 1 ml D_2O (containing 0.05 % w/v TSP) and centrifuged at 13 000 rpm for 15 minutes. A sample volume of 800 µl was then taken for NMR analysis.

For analysis using organic extraction, chloroform was substituted for distilled water, followed by evaporation using nitrogen gas. Samples were reconstituted in d_4-methanol. Other conditions were as for the aqueous extraction.

HPLC analysis

All samples were dissolved at a concentration of 10 mg ml^{-1} in acetonitrile and filtered through 0.2 mm PTFE filters. HPLC analysis was carried out for each sample using the following procedure: 20 µl of each sample were injected onto a 5 µ ODS reverse phase column (Hypersil 250 × 4.6 mm) and compounds were separated using the solvent system shown below (flow rate 1 ml/min). Compounds were detected at a wavelength of 210 nm. A calibration curve was constructed using parthenolide standards.

Time (min)	*% acetonitrile*	*% water*
0.00	10	90
4.00	10	90
5.00	40	60
30.00	40	60

^{1}H-NMR spectroscopy

NMR spectra were run on a Bruker (Bruker GmbH, Rheinstetten, Germany) DRX 600 Spectrometer, operating at 600.13 MHz for

the ^{1}H frequency, utilising Bruker BEST flow injection technology for sample transfer. Spectra were the result of the summation of 64 free induction decays (FIDs), with data collected into 48 k datapoints, and a sweep width of 20.03 ppm. Acquisition time was 2.04 seconds. The water signal was suppressed using a standard 1D-pulse sequence [18]. Prior to Fourier transformation, an exponential line broadening equivalent to 0.3 Hz was applied to the FIDs. Spectra were referenced to TSP at 0.00 ppm.

Principal Components Analysis (PCA)

NMR spectra were reduced to 252 regions by digitisation to produce a series of sequentially integrated regions δ = 0.04 in width between δ = 0.06 and 9.98, using Bruker AMIX software (version 2.0, Bruker GmbH, Germany). The resulting data were exported into Microsoft® Excel in the form of a bar chart. Following the removal of the regions not related to the signals of interest, i.e., around the residual water signal (δ = 4.54 to 4.98) and TSP (δ = −0.02 to 0.02), 237 integral regions remained. The regions were normalised to the whole spectrum for subsequent PCA.

PCA was performed using SIMCA-P 8.0 multivariate data analysis software (Umetrics, Sweden), with mean centring of the data preceding PCA. The output from the PCA analysis consisted of scores plots (giving an indication of the separation of the classes in terms of chemical similarity), and loadings plots, which give an indication as to which NMR spectral regions were important with respect to the classification obtained in the scores plots.

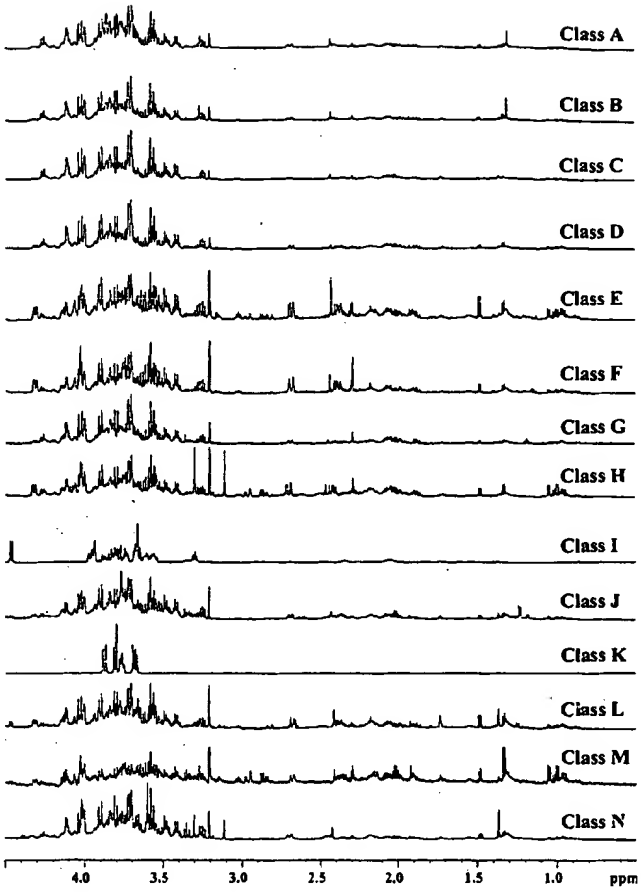


Fig. 1 600 MHz ^{1}H-NMR spectra for all 14 feverfew sample classes.

Results and Discussion

Initial observations of ^{1}H-NMR spectral data

The ^{1}H-NMR spectra of the feverfew samples from the 14 different suppliers (A-N) analysed can be seen in Fig. 1. It can be seen from this figure that the spectra are reasonably similar. Obvious exceptions to this are class I and class K which appear very different from the remaining twelve classes. These remaining classes, however, are all of similar appearance, although differences are discernible under close scrutiny. Because of this broad similarity of the sample classes, it is necessary to use a multivariate data analysis approach to interrogate the data to a greater extent than is possible by direct observation of the ^{1}H-NMR spectrum alone.

PCA analysis of ^{1}H-NMR data

Principal components analysis (PCA) following multivariate analysis of the ^{1}H NMR spectra for 14 different feverfew samples (classes) showed that while two classes (I and K) were well separated with respect to the other samples in the first three PCs (PC1-PC3), the remaining twelve classes were difficult to differentiate [Fig. 2 (a)]. This indicated that the two classes that were well separated spectroscopically, and hence were chemically very different to the remaining twelve classes, and so the variance in the first two PCs was essentially due to differences between the two chemically different classes and the remaining twelve classes. Thus these two classes have a high leverage on the overall model, which becomes skewed.

As indicated above, the ^{1}H-NMR spectra for classes I and K (Fig. 1) when compared with the other twelve classes are in fact extremely different. Indeed, the spectrum for class K in particular resembles that of a single component (tentatively identified as mannitol) rather than a multi-component extract. Because of this large difference between samples from classes I and K and the remaining twelve classes, it was appropriate to exclude these two classes from the analysis, and to carry out a subsequent PCA on the remaining twelve classes. The 3D-plot of PC1-PC3 following the PCA of the dataset containing twelve of the classes is shown in Fig. 2 (b). It can be seen that the clustering of these twelve classes into discrete groups is much more apparent following the removal of two classes I and K, respectively. A combination of the first 3 PCs, accounting for 84.5% of the total variance allows separation of all the remaining twelve classes.

In order to ensure that clustering was not as a result of different excipients present in some of the samples obtained in tablet form, the NMR analysis was repeated using chloroform extracts (thus obtaining an organic extract that would not contain ingredients such as glucose). Similar clustering to that obtained in aqueous extracts was obtained from the organic extracts demonstrating the robustness of the technique (Fig. 3).

Comparison of ^{1}H-NMR and HPLC-UV data for PCA classification

PCA analysis of feverfew has been performed previously using HPLC-UV data [8]. We compared the performance of PCA of HPLC-UV and ^{1}H-NMR of extracts on five selected sample classes (A, B, F, I and L) to evaluate the classification potential of both techniques. The scores plot obtained after performing PC analysis on HPLC-UV data from feverfew extracts (using a previously published HPLC method [19]) from the five selected classes is

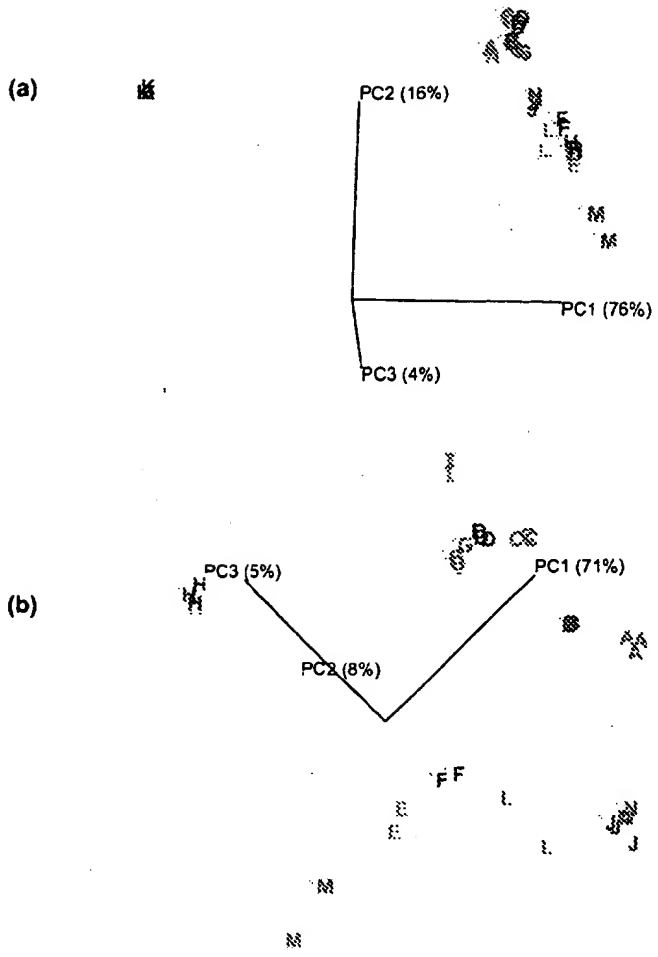


Fig. 2 3D PCA plot for PC1-PC3 for (a) 14 commercially available samples of feverfew and (b) for twelve commercially available samples of feverfew, with two outlying classes (I and K) removed from the previous analysis following NMR spectroscopic analysis of aqueous extracts.

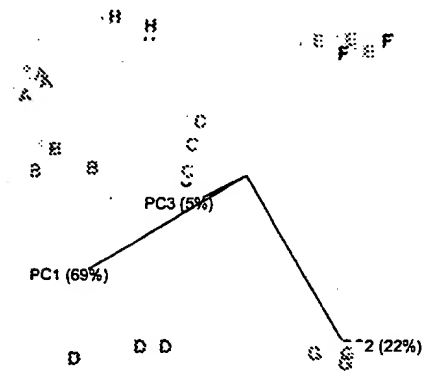


Fig. 3 3D PCA plot for PC1-PC3 for 8 commercially available samples of feverfew following NMR spectroscopic analysis of organic extracts.

shown in Fig. 4 (a) and shows some separation between the five classes analysed. However, when compared to the PC plot obtained after analysis of the ^{1}H-NMR data on the whole extract [Fig. 4 (b)], it is apparent that the NMR analysis results in better separation and tighter clustering of data (in particular, the scale on the axes is 2 orders of magnitude smaller in the NMR plot). The average standard deviations for each PC plot (obtained by averaging the standard deviations for each class within each plot), demonstrate this quantitatively, with the standard deviations for the HPLC plot being 369 and 172 for PC1 and PC2, respectively, while the NMR plot has standard deviations of 0.1 for both PC1 and PC2. Furthermore, it can be seen that while the NMR data clearly separates the outlying class I from the other four classes in PC1 [Fig. 4 (b)], the HPLC analysis [Fig. 4 (a)] results in a less clear distinction between a class that is clearly very different from the others based on ^{1}H-NMR data. The differences in discriminating ability are due to the lack of dependence of NMR spectroscopy on differential chromophoric strength of separated metabolites and the representative distribution of metabolite concentration based on ^{1}H-NMR signal intensity. The NMR data also have an intrinsically higher dimensionality (information content) than the HPLC-UV data. While HPLC requires a chromophore at a particular wavelength in order to detect a particular compound, NMR will detect all ^{1}H-containing species in a sample. Therefore, although the HPLC chromatograms from the five classes were seen to be broadly similar (data not shown), interrogation of the NMR spectra indicates clear differences between the classes, as discussed above (Fig. 1). It is this ability to detect a wide range of components within a complex mixture that makes NMR a powerful technique for such analyses.

A further sample set was used to investigate whether PCA could distinguish between different batches of samples from the same supplier. The scores plot for PC1/PC2 for two classes (labelled W and X) of sample from the same supplier, but with different batch numbers is shown in Fig. 5 (a). It can be clearly seen that even with these samples that should have very similar composition, inter-batch variation is easily identified. When two different bottles of tablets from the same supplier carrying the same batch number are studied, however [labelled Y and Z, Fig. 5 (b)],

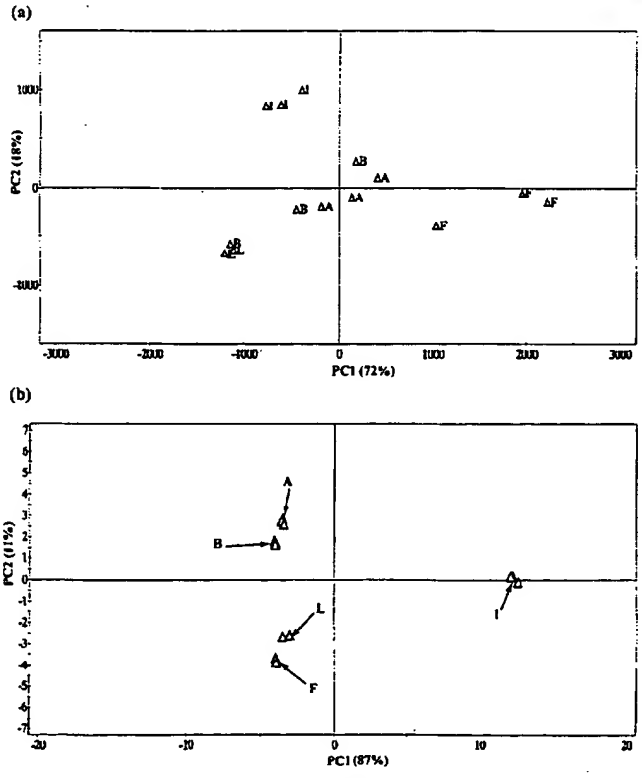


Fig. 4 PCA plots of five commercially available samples of feverfew following (a) HPLC and (b) NMR data acquisition.

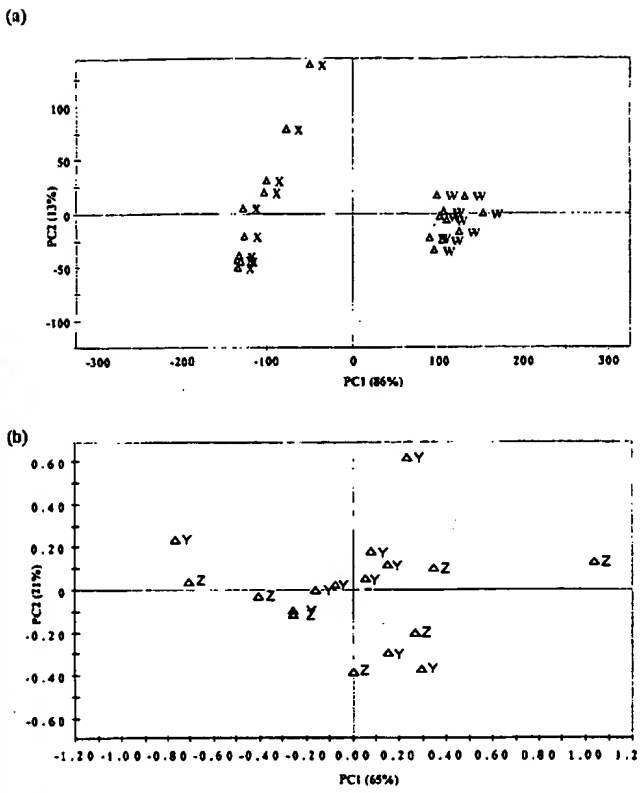


Fig. 5 PCA plots (a) PC1 v PC2 for two classes of sample obtained from the same supplier, but with two different batch numbers and (b) PC2 v PC4 for two classes of sample obtained from the same supplier, with the same batch number.

it was noted that slight separation was achieved although the two classes were less readily distinguished than in the previous example. These data suggest that although inter-batch variation is detectable using the combination of NMR and PCA, intra-batch variation, as would be expected, is less easily observed.

Conclusions

The results presented here show that it is possible to use ^{1}H-NMR spectroscopy and multivariate data analysis to discriminate between feverfew samples from different suppliers. In particular, it has been shown that samples that really are very different from the 'average' feverfew sample are easily identified when compared to other samples. This demonstrates one of the major advantages of NMR spectroscopy for direct multi-component analysis over other analytical techniques, such as HPLC, in that unexpected results such as this are easily observed due to the non-selective nature of NMR spectroscopic experiment acquisition. In addition, it has been shown that the use of multivariate data analysis can readily discriminate between very similar samples of feverfew extracts.

References

[1] Johnson E, Kadam N, Hylands D, Hylands P. Efficacy of feverfew as prophylactic treatment of migraine. Brit Med J 1985; 291: 569–73

[2] Awang DV, Dawson BA, Kindack DG, Crompton CW, Heptinstall S. Parthenolide content of feverfew (*Tanacetum parthenium*) assessed by HPLC an ^{1}H-NMR spectroscopy. J Nat Prod 1991; 54: 1516–21

[3] Awang DV, Dawson BA, Kindack DG, Crompton CW, Heptinstall S. In 30th Annual Meeting of the ASP. San Juan, Puerto Rico: 1989

[4] Heptinstall S, Awang DV, Dawson BA, Kindack DG, Knight D, May J. Parthenolide content and bioactivity of feverfew (*Tanacetum parthenium* (L.) Schultz-Bip.). Estimation of commercial and authenticated feverfew products. J Pharm Pharmacol 1992; 44: 391–5

[5] Jessup D. PhD. University of London, 1982

[6] Williamson E. Synergy – fact or fiction? In Herbal medicine: A concise overview for professionals. E. Ernst. Butterworth-Heinemann, Oxford: 2000: 43–58

[7] Jain AK, Duin RPW, Mao J. Statistical pattern recognition: A review. IEEE T Pattern Anal 2000; 20: 4–37

[8] Smith RM, Burford MD. Comparison of flavonoids in feverfew varieties and related species by principal components analysis. Chemometr Intell Lab 1993; 18: 285–91

[9] Lazarowych NJ, Pekos P. Use of fingerprinting and marker compounds for identification and standardisation of botanical drugs: Strategies for applying pharmaceutical HPLC analysis to herbal products. Drug Inf J 1998; 32: 497–512

[10] Knight D. Feverfew: Chemistry and biological activity. Nat Prod Rep 1995; 12: 271–6

[11] Williams CA, Harborne JB, Geiger H, Hoult JRS. The flavonoids of *Tanacetum parthenium* and *T. vulgare* and their anti-inflammatory properties. Phytochemistry 1999; 51: 417–23

[12] Nicholson J, Lindon J, Holmes E. Metabonomics: Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 1999; 29: 1181–9

[13] Holmes E, Shockcor JP. Accelerated toxicity screening using NMR and pattern recognition-based methods. Curr Opinion Drug Disc Dev 2000; 3: 72–8

[14] Belton PS, Colquhoun IJ, Kemsley EK, Delgadillo I, Roma P, Dennis MJ. Application of chemometrics to the ^{1}H-NMR spectra of apple juices: Discrimination between apple varieties. Food Chem 1998; 61: 207–13

[15] Forveille L, Vercauteren J, Rutledge DN. Multivariate statistical analysis of two dimensional NMR data to differentiate grapevine cultivars and clones. Food Chem. 1996; 57: 441–50

[16] Alam TM, Alam MK. Chemometric analysis of NMR spectroscopy data. Spectroscopy 2001; 16: 18–25

[17] El-Deredy W. Pattern recognition approaches in biomedical and clinical magnetic resonance spectroscopy: A review. NMR Biomed 1997; 10: 99–124

[18] Nicholson J, Foxall PJ, Spraul M, Farrant RD, Lindon J. 750 MHz ^{1}H- and ^{1}H-^{13}C-NMR spectroscopy of human blood plasma. Anal Chem 1995; 67: 793–811

[19] Brown AM, Lowe KC, Davey MR, Power JB, Knight D, Heptinstall S. Comparison of extraction procedures for parthenolide in *Tanacetum parthenium*. Phytochem Anal 1996; 7: 86–91

Mellerson, Kendra

From: Gakh, Yelena
Sent: Tuesday, August 05, 2003 2:33 PM
To: STIC-EIC1700
Subject: 09890973

Dear Kendra:

please order one more list:

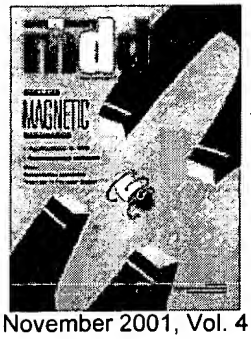
7. TITLE: "Screening with NMR"
AUTHOR(S): *Bradley, David*
CORPORATE SOURCE: Washington DC, USA
SOURCE: **Modern Drug Discovery (2001), 4(11), 28-30,32,34**

Thank you,

Yelena

Yelena G. Gakh, Ph.D.

Patent Examiner
USPTO, cp3/7B-08
(703)306-5906



November 2001, Vol. 4
No. 11, pp 28–30, 32, 34.

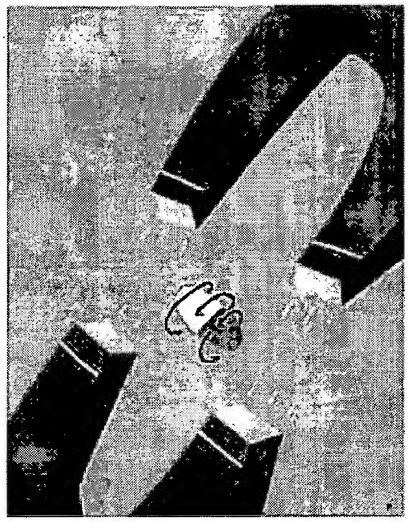
modern drug discovery
mdd
from concept to development

Focus: High Throughput / Robotics
Feature Article

Screening with NMR

DAVID BRADLEY

Advances in NMR automation have allowed researchers to follow drug development from beginning to end.



Nuclear magnetic resonance (NMR) spectroscopy has been widely adopted since its invention. What was once a cumbersome technique can now reveal the most cryptic details of sophisticated molecular systems. It is one of the most information-rich analytical techniques. The latest machines can place very small samples in a magnetic field with a strength of more than 21 T and detect radio-frequency signals of almost 1 GHz. Systems capable of automated and high-throughput sampling are poised to push NMR into the mainstream, not just as the analytical tool of choice but as a key component of the drug discovery process.

NMR speeds up

Several research teams are working on bringing NMR spectrometers into drug discovery laboratories and using them to further accelerate the rate of pharmaceutical R&D. According to researchers at Varian (Palo Alto, CA), one of the serious drawbacks in getting the best results from a combinatorial array

is the inability to obtain a complete sample analysis.

In pioneering work on LC-NMR carried out by Jeremy Nicholson and John Lindon at Imperial College (London), in collaboration with Manfred Spraul of Bruker GmbH, Nicholson's team separated and assigned a randomly synthesized collection of 27 tripeptides—all the combinations of Ala, Tyr, and Met—using one chromatographic run that took about 30 min (*1*). In Nicholson's words, "Not a bad first attempt!" Varian scientists recently extended Nicholson's research to other areas of combinatorial chemistry by devising an automated approach to NMR that allows combinatorial chemists to quickly and easily obtain the ^{1}H-NMR spectra of solution-phase samples.

The Varian team worked on obtaining the NMR spectra of compounds bound to solid supports and was rewarded with the rapid adoption of its techniques throughout the combinatorial community. Unfortunately, the teams' solid-state NMR approach is confined to analyzing small numbers of samples and lacks the high-throughput capability needed for efficient analysis of vast compound libraries. A flow technique coupled with automated sample analysis using liquid-phase NMR would help the analyst rein in combinatorial libraries.

While developing HPLC-NMR techniques, the Varian team realized that the LC-NMR approach could be refined as a useful tool for combinatorial applications. Combinatorial chemistry not only traditionally generates large numbers of compounds in small quantities, but also tends to do away with the use of conventional glassware, replacing it with the increasingly familiar multiple-welled microtiter plates. To address these issues, Varian scientists built and tested a flow-NMR sample changer. "The system reduces the cost, time, and effort of sample handling, allows inexpensive sample containers to be used, and uses smaller quantities of sample than traditional automated NMR systems," according to Varian.

The flow-NMR approach precludes the need for transferring samples from the microtiter plates to NMR tubes, which would be the biggest cost in high-resolution NMR of a large library, for which not only precision glass tubes and deuterated solvents are required for each sample from each cell, but also a drying (solvent removal) process. Instead, the team at Varian used an automated liquid-handling device, such as the Gilson Model 215

Liquids Handler, which takes a sample solution stored in a microtiter plate and injects it directly into an NMR flow probe.

With each step of the protocol controlled by a computer, the system first rinses the NMR flow cell with a solvent and disposes the waste solvent. The liquid handler then moves a controlled volume of the appropriate sample into the NMR probe, at which point the spectrometer is signaled to begin gathering data. The process can be repeated automatically with any number of NMR experiments on each sample. The team refers to the approach as direct injection (DI) NMR; and the liquid handler is referred to as the versatile automated sample transport (VAST).

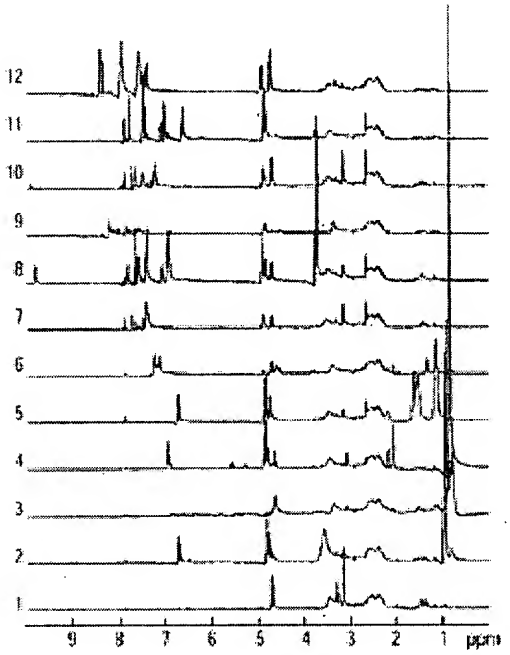


Figure 1. Seeing how things develop. Using automated systems developed by companies such as Bruker and Varian, researchers can quickly generate ^{1}H-NMR spectra of compounds synthesized in a 96-well plate. (Adapted from Reference *2*.)

The DI VAST approach can quickly gather one-dimensional ^{1}H-NMR spectra for each member of a combinatorial library, an approach that the team says is almost routine at Varian and elsewhere (Figure 1). For example, at Monsanto (St. Louis) Bruce Hamper and his team used the VAST system to characterize a 96-member substituted methylene malonamic acid library (*2*).

"This only works in libraries that have one compound per well,"

points out Lenore Martin, assistant professor in the department of biochemistry, microbiology, and molecular genetics at the University of Rhode Island. The standard in the industry is to have groups of compounds in each well, so there is still a requirement to couple the flow cell to a separation technique such as LC. "Another very promising technique is capillary electrophoresis (CE)-NMR," adds Martin, "which is being developed by a group in the department of chemistry at the University of Illinois, Champaign-Urbana."

Drug design by NMR

NMR is ideal for screening fragments of potential drug molecules, according to the work of Stephen Fesik of Abbott Laboratories (Abbott Park, IL). Recently, he and his colleagues devised a strategy for designing high-affinity ligands to create drugs that inhibit kinases (*3*). Fesik says that finding leads of sufficient specificity, bioavailability, and safety is "still an arduous process" and usually has a failure rate of 50% in the initial stages of drug discovery. A method to bump up successes without added synthetic effort would be useful. Fesik's "fragment" approach fits the bill and involves screening a range of fragments that could be incorporated into an inhibitor without reducing potency but improving characteristics, such as solubility or reduced toxicity.

The first step is to fragment an existing lead molecule, identify a range of suitable replacements for the fragments, and build these into the original molecular skeleton. The problems arise in trying to identify suitable fragments. The fragments bind weakly to the target receptor or enzyme, so conventional screening methods cannot reliably detect their binding, because high concentrations are required to generate a detectable response. Moreover, standard assays indicate nothing about binding orientation or site, and so they offer no clues about optimal positioning of the fragment on the skeleton.

Fesik and his colleagues found a way to screen such fragments successfully by using NMR based on a Bruker system. The affinity and binding site location of the chosen fragment are determined by watching how the ^{15}N–^{1}H heteronuclear single quantum coherence (HSQC) spectra of the ^{15}N-labeled protein change when the test molecule is added. The next step involves using NMR to identify molecules that bind to the same site as the chosen fragment. The fragments identified can then be

incorporated into the skeleton for further study.

"This approach is a valuable strategy for modifying existing leads to improve their potency, bioavailability, or toxicity profile, and thus represents a useful technique for lead optimization," says Fesik. Moreover, he emphasizes that the use of NMR in this manner means that thousands of potential mimetics with a range of functionality can be quickly analyzed without the need for multiple synthetic routes to be implemented and thousands of putative leads prepared. Indeed, the Fesik team previously demonstrated high-throughput NMR that could investigate potential ligands for unknown proteins at a rate of 200,000 per month (*4*).

Toward proteomics

If NMR is going to respond to the postgenomic challenge of addressing thousands of new drug targets, innovations are needed to remove two key limitations. First, NMR structural studies cannot be performed for proteins much larger than 35 kD. Second, to attack thousands of proteins, a proteomically leveraged, highly parallel strategy to drug design is needed; but current strategies attack one target at a time. Triad Therapeutics in San Diego is removing both of these barriers, thus extending NMR drug discovery efforts in a proteome-wide manner.

Triad developed a suite of NMR technologies that allow for the characterization of protein–ligand interactions with unprecedented speed (days as opposed to months). These tools, combined with bioinformatics strategies, allow the systematic gathering of information that describes protein–ligand interactions across large gene families of proteins such as kinases and dehydrogenases. The term "enzyme mechanomics" describes this newly enabled gene-family-wide characterization of structure–function correlations.

"Triad makes use of a technology called NMR SOLVE—structurally oriented library valency engineering—to guide the design of combinatorial libraries tailored to entire gene families of proteins, using the enzyme mechanomic data," says Daniel Sem, Triad's vice president of biophysics. He and colleague Maurizio Pellecchia point out that NMR is intrinsically a noninvasive technique and thus is ideally suited to observing the dynamics of a molecular system, as well as acting as an

analytical tool.

"Any NMR method that provides structural information on large proteins must provide a way to simplify NMR spectra—to focus in on that part of a spectrum corresponding to atoms that are in a protein's binding site," explains Sem. As such, Sem, Pellecchia, and colleagues at the University of Wisconsin have devised a technique that can reduce overlap in protein spectra and allow these complex biomolecules to be investigated in their native state with much greater clarity (*5*). This method, called solvent-exposed amides with transverse relaxation-optimized spectroscopy (SEA-TROSY), is combined with other experiments to look at very large protein structures, their backbone dynamics, and how ligands or inhibitors bind to them.

"NMR is now poised to tackle the postgenomic challenge of attacking large numbers of new drug targets with greater speed, in a highly parallel manner, and without the usual limitation to low-molecular-weight proteins," adds Sem.

The metabolic end point

One approach to drug research closely considers the end product of the drug cycle. Jeremy Nicholson uses high-resolution NMR to screen body fluids and magic-angle spinning NMR to screen tissues for metabolic byproducts of drugs and to detect perturbations in endogenous metabolic profiles in disease processes (*6, 7*). Nicholson and his colleagues have spent the past two decades looking into metabonomics, a field driven mainly by NMR spectroscopy. Nicholson describes metabonomics, a term he coined about six years ago, as the "quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification."

Rather than focusing on single analytes as might be the case in a clinical diagnostics approach, Nicholson's team has used ^{1}H-NMR to build up expertise in the multicomponent metabolic composition of cells, tissues, and biological fluids (saliva, blood, urine, semen, and even sweat). The team uses pattern recognition, expert systems, and related bioinformatic tools to interpret and classify the complex data sets generated by one- and two-dimensional NMR analysis of such samples. They can now spot telltale metabolic fingerprints in NMR spectra. NMR, in particular, gives a very complex fingerprint of a large number of

metabolite signatures—thousands in the case of a urine sample (Figure 2).

"The quantitative analysis of such profiles gives insight into sites and mechanisms of toxicity according to the characteristic perturbations in the metabolic profile," explains Nicholson. "Biomarker information can be statistically extracted from spectra and, as NMR is a structural organic chemistry tool, novel metabolic markers can be structurally characterized.

"The recovery of high-density metabolic information from complex spectra is facilitated by the use of an array of multivariate statistical and pattern recognition tools that classify toxicity or disease state according to spectral profile and identify critical regions of the NMR spectral fingerprints that are modified by the pathological process," says Nicholson. Exact biomarker identification is then achieved or confirmed by judicious use of multidimensional NMR spectroscopy (e.g., ^{1}H-^{13}C HSQC or heteronuclear multiple-bond correlation spectroscopy) combined with HPLC–NMR–mass spectrometry methods (*8*).

A holistic picture

The London team also recently introduced the concept of "integrated metabonomics". This, Nicholson says, is the parallel NMR investigation of multiple biological fluids, and sometimes selected tissue samples, using magic-angle spinning NMR methods at various time points after drug exposure to gain a holistic picture of a series of metabolic events in the whole body.

Nicholson and his colleagues are now involved in cross-correlating integrated metabonomic data with those generated by genomics and proteomics (what he terms "integrated bionomics") to describe the biochemical consequences of pathological processes at multiple levels of biomolecular organization and to learn about silent gene function.

From humble beginnings as a simple spectroscopic tool for working out molecular structures, NMR has raced to the front of the drug discovery arsenal, providing pharma researchers with a powerful weapon with which to hack through the molecular jungle.

References

1. Lindon, J. C.; et al. *Magn. Reson. Chem.* **1995,** *33,* 857–863.
2. Hamper, B. C.; et al. *J. Comb. Chem.* **1999,** *1,* 140–150.
3. Hadjuk, P. J.; et al. *J. Med. Chem.* **2000,** *43,* 4781–4786.
4. Hadjuk, P. J.; et al. *J. Med. Chem.* **1999,** *42,* 2315–2317.
5. Pellechia, M.; et al. *J. Am. Chem. Soc.* **2001,** *123,* 4633–4634.
6. Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica* **1999,** *29,* 1181–1189.
7. Lindon, J. C.; et al. *Concepts Magn. Reson.* **2000,** *12,* 289–320.
8. Lindon, J. C.; Holmes, E.; Nicholson, J. K. *Prog. Nucl. Magn. Reson. Spectrosc.* **2001,** *39,* 1–40.
9. Holmes, E.; et al. *Chem. Res. Toxicol.* **2000,** *13,* 471–478.

David Bradley *is a freelance writer living in Cambridge, UK. Send your comments or questions regarding this article to* mdd@acs.org *or the Editorial Office by fax at 202-776-8166 or by post at 1155 16th Street, NW; Washington, DC 20036.*

Return to Top || Table of Contents

RM 301.25
~~HL 10~~, C87
NOS

STIC-ILL

458449

From: Mellerson, Kendra
Sent: Tuesday, August 05, 2003 4:53 PM
To: STIC-ILL
Subject: FW: 09890973

-----Original Message-----
From: Gakh, Yelena
Sent: Tuesday, August 05, 2003 2:33 PM
To: STIC-EIC1700
Subject: 09890973

Dear Kendra:

please order one more list:

8. TITLE: "Accelerated toxicity screening using NMR and pattern recognition-based methods"
AUTHOR(S): *Holmes, Elaine; Shockcor, John P.*
CORPORATE SOURCE: Department of Biological Chemistry, Division of Biomedical Sciences, Imperial College of Science, Technology and Medicine, London, SW7 2AZ, UK
SOURCE: **Current Opinion in Drug Discovery & Development (2000), 3(1), 72-78**

Thank you,

Yelena

Yelena G. Gakh, Ph.D.

Patent Examiner
USPTO, cp3/7B-08
(703)306-5906

CBS - NO
eyn 8/7 - NO (call)
BRI 8/7/03

Accelerated toxicity screening using NMR and pattern recognition-based methods

Elaine Holmes[1] & John P Shockcor[2]

Addresses
[1]Department of Biological Chemistry
Division of Biomedical Sciences
Imperial College of Science, Technology and Medicine
Sir Alexander Fleming Building
Exhibition Road
London SW7 2AZ
UK
Email: elaine.holmes@ic.ac.uk

[2]DuPont Pharmaceuticals Company
Stine-Haskell Research Center
Bldg 115
PO Box 30
Elkton Road
Newark
DE 19714
USA
Email: john.p.shockcor@dupontpharma.com

Current Opinion in Drug Discovery & Development 2000 3(1):72-78
 ISSN 1367-6733

^{1}H-NMR spectroscopy has proved to be a powerful and efficient means of monitoring the interaction of pharmacological agents with cells and tissues [1•]. The application of this technique to biofluid analysis, gives rise to a comprehensive metabolic profile of the low molecular weight components of biofluids, that reflect concentrations and fluxes of endogenous metabolites involved in key internediary cellular pathways, thereby giving an indication of an organism's physiological or pathophysiological status [1•]. Recent developments in spectrometer technology have resulted in increased sensitivity and dispersion. Together with the increased capacity for sample throughput (~ 300 samples/day), arising from the latest advances in flow probe technology and in robotic transfer systems [2], ^{1}H-NMR spectroscopic techniques have become viable in terms of toxicological screening. However, the complexity of high-field biofluid spectra in conjunction with the increased capacity for sample handling, leading to a rapid growth in the size of toxicological spectral databases, has placed greater emphasis on the need to develop improved automated procedures for data processing and interpretation. By harnessing chemometric tools to the analysis of complex spectral data, the toxicological consequences of xenobiotic exposure can be evaluated efficiently on-line. Automation of spectral processing procedures and the construction of mathematically-based 'expert systems' for the prediction of drug-induced toxicity founded on ^{1}H-NMR spectral profiles, have now been achieved. In this article, we review the recent developments in NMR and pattern recognition analysis and consider their application in toxicological screening.

Keywords Biomarker, ^{1}H-NMR spectroscopy, metabolic profile, metabonomics, pattern recognition, toxicological screening

Abbreviations

BEA	2-bromoethylamine
DMG	*N,N*-dimethylglycine
HCBD	hexachlorobutadiene
MAS	magic angle spinning
pd	post-dose
SD	Sprague-Dawley
TMAO	trimethylamine-*N*-oxide

Introduction

The current emphasis in the pharmaceutical industry placed on optimizing the efficiency of lead compound selection and minimizing overall attrition rates, has led to the extensive evaluation of new analytical technologies such as proteomics and genomics. Whilst genomics allows the measurements of responses of living systems to drugs at the genetic level, and proteomics enables the response of an organism at the level of cellular proteins to be assessed [3,4], neither technology provides an holistic picture of a toxicological episode. In order to understand fully the pathophysiological processes induced by xenobiotics, the metabolic status of the whole organism needs to be taken into account. Metabonomics, defined as 'The quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification', provides an efficient means of measuring the metabolic response of an organism to xenobiotic exposure [5••], and is complementary to any information obtained from genomic and proteomic analysis. The concept of metabonomics has evolved from the work of Nicholson and co-workers and is founded on two decades of ^{1}H-NMR spectroscopic analysis of the multi-component metabolic composition of biofluids, cells and tissues under different physiological and pathophysiological conditions [6-14]. In this review, we summarize the major events in the evolution of NMR-based metabonomics and discuss the application of this technique as a toxicological probe, both for characterizing site or mechanism-specific toxicity and for identifying toxicological biomarkers *in vivo*.

Background to the application of ^{1}H-NMR spectroscopy in toxicology

High-resolution ^{1}H-NMR spectroscopic analysis of biofluids has proved to be one of the most powerful techniques for investigating the response of organisms to xenobiotics. Comprehensive profiles of metabolite signals can be obtained without the need for preselection of measurement parameters or selection derivatization procedures [1•,6-13]. Furthermore, bioanalytically, ^{1}H-NMR spectroscopic analysis of biofluids is more efficient than the methods used to characterize either the genetic or proteomic composition of samples. Analysis is non-destructive, cost-effective and typically takes only a few minutes per sample, requiring little or no sample pre-treatment or reagents.

Exposure of an organism to a xenobiotic, results in subtle modifications in the biochemical composition of intra- and extracellular fluids as the organism attempts to maintain homeostasis (constancy of internal environment). This adjustment results in alterations in the composition of body fluids, such as urine and plasma, which can be profiled using ^{1}H-NMR spectroscopic analysis. The ^{1}H-NMR spectral profiles of biofluids provide a 'unique' fingerprint of the metabolic state of an organism and can provide information on the nature of drug or toxin to which an animal has been exposed [1•,5••]. Characteristic changes in the

concentrations and patterns of endogenous metabolites in biofluids are often indicative of the site or basic mechanism of toxicity. For example, increased urinary levels of glucose, organic and amino acids are indicative of damage to the S_2 segment of the renal cortex [6], whilst increased urinary excretion of taurine and creatine generally reflect a hepatotoxic lesion [15•]. ^{1}H-NMR spectral profiles of urine obtained after treating rats with various model nephrotoxins that target specific regions of the kidney are shown in Figure 1. Each nephrotoxin produces a characteristic spectral profile, with compounds that target the same region, eg, $HgCl_4$ and hexachlorobutadiene (HCBD), giving rise to similar metabolic profiles (although the profiles of any compounds will be unique). The biochemical consequences of over 100 drugs and model toxins have been characterized metabonomically via ^{1}H-NMR spectroscopy of biofluids, such as urine, plasma and bile, and large spectral databases describing toxicological events have been constructed [5••]. However, in reality, toxicological data are exceptionally complex. Lesions develop and resolve in real time and hence, time-related changes in NMR-detected metabolic profiles for each toxin must be taken into account, and indeed the time profile itself is a feature of the toxicity [16•,17] (Figure 2). In addition, drugs rarely specifically target a single organ and most will inevitably induce biochemical effects in a range of tissues. Therefore, ^{1}H-NMR spectra of biofluids represent complex indices of the metabolic response of an organism to xenobiotic exposure. However, despite the inherent complexity of high-field ^{1}H-NMR biofluid spectra, numerous novel metabolic biomarkers of organ-specific toxicity in the rat have been successfully elucidated. For example, renal papillary necrosis was a condition for which no early biochemical markers of damage previously existed. However, following ^{1}H-NMR spectroscopic analysis of urine obtained from rats treated with model renal papillary toxins, perturbations in the levels of trimethylamine-*N*-oxide (TMAO), *N,N*-dimethylglycine (DMG), dimethylamine and succinate were found to be indicative of damage to the renal papilla [6,18]. However, the biomarker information within NMR spectra of biofluids is much more subtle and rich than a small set of biochemicals defining a single metabolic event. Hundreds of compounds representing many pathways can often be measured simultaneously, and it is the overall metabonomic response to toxic insult (occurring over time), that characterizes a lesion so well [16•,19].

Figure 1. Stackplot of 600 MHz ^{1}H-NMR spectra of urine obtained from rats treated with a range of tissue-specific nephrotoxins showing characteristic patterns of metabolic perturbation.

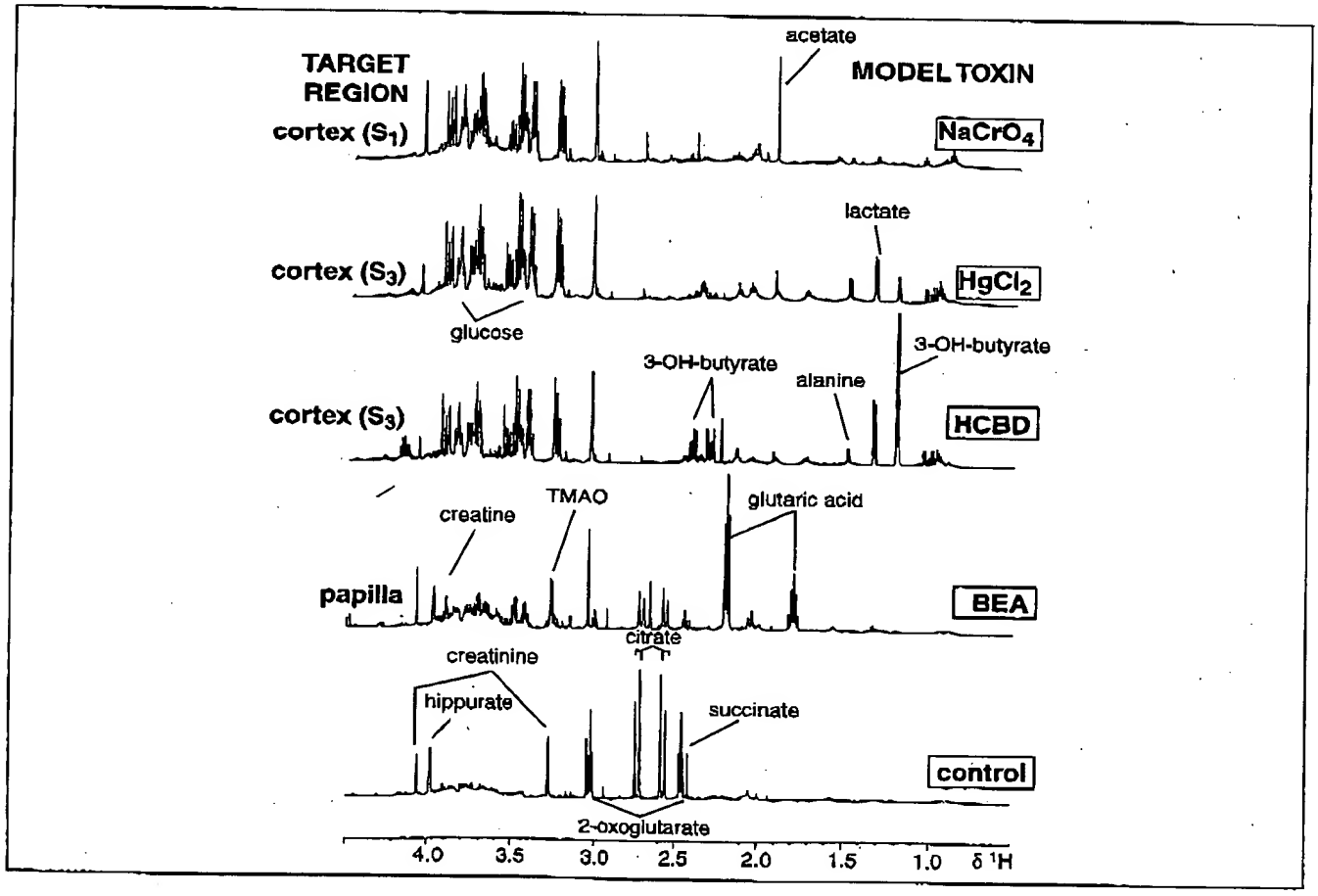
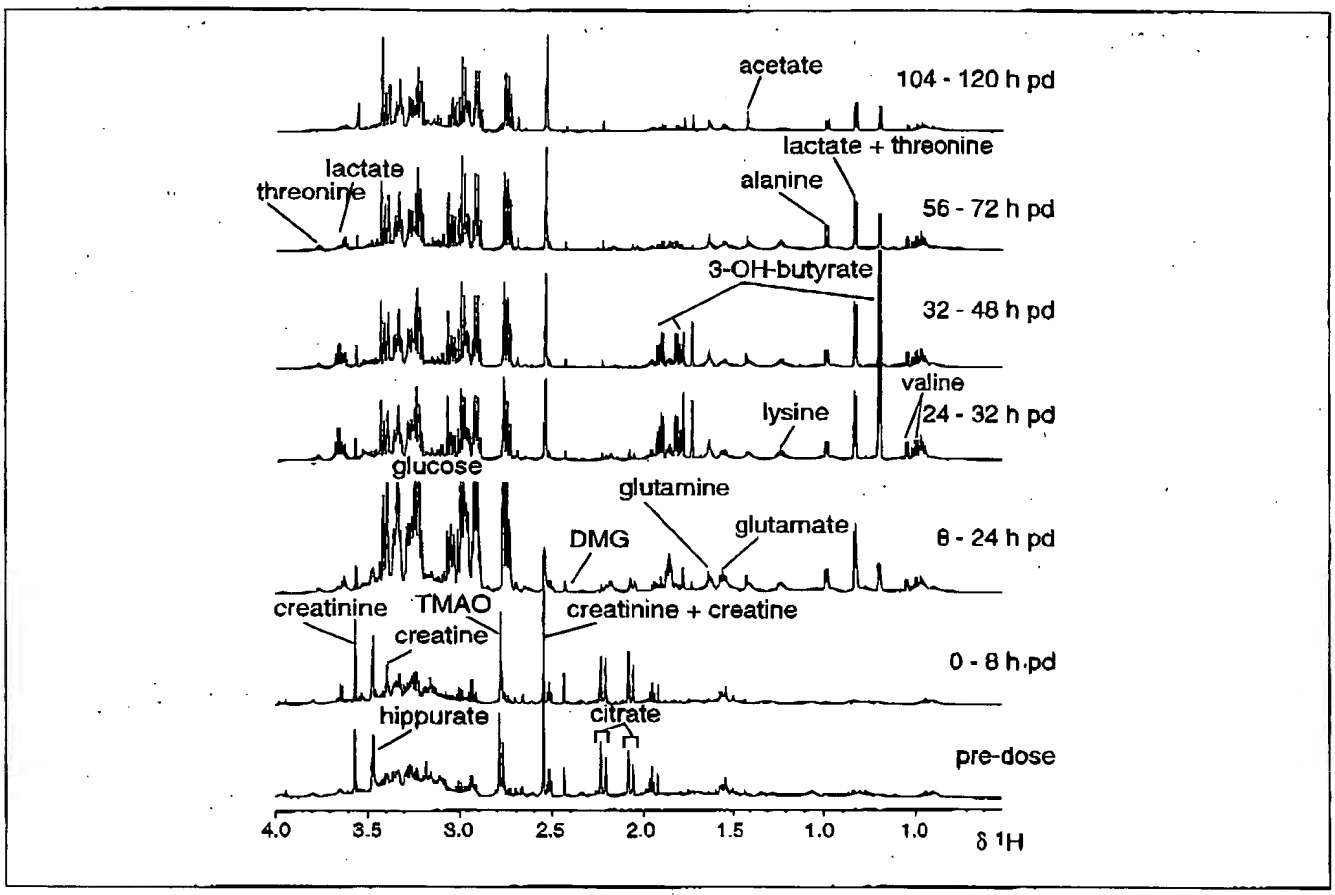


Figure 2. Stackplot of 600 MHz ^{1}H-NMR spectra of urine obtained over a 7-day period following the administration of HCBD (200 mg/kg) to a SD rat.



pd = post-dose

Recent advances in NMR spectroscopy of biofluids and tissues

Recent improvements in instrumentation and the increasing availability of ultra-high-frequency NMR spectrometers have resulted in increased sensitivity and dispersion, which increases the amount of latent metabolic information in the biofluid spectra [19]. Typically, 600 to 800 MHz ^{1}H-NMR spectra of biofluids, such as urine and plasma, contain thousands of signals arising from hundreds of endogenous molecules representing many biochemical pathways [5••,19]. Although the potential of ^{1}H-NMR spectroscopy for classifying toxic lesions and for elucidating markers of toxicity increases with field strength, the complexity of these spectra generally requires the use of data reduction and pattern recognition (PR) techniques in order to access the latent biochemical information present in the spectra. Multivariate analyses of NMR automated data reduction procedures have been used to remove subjective bias in the choice of spectral descriptors and to improve the efficiency of the analysis. Automated flow and robotic systems have increased the capacity for data accumulation and resulted in a backlog of data processing and analysis. Recent developments in software packages enabling automatic phase correction, referencing and data reduction accommodate this increased sample throughput. Moreover, software packages capable of performing multivariate statistical analysis on spectral data are also commonly available and provide a means of data reduction and visualization in order to explore intrinsic toxin-related clustering behavior of samples.

With the advent of ultra-high-field spectrometers, perturbation in the levels of metabolites present at low concentrations can be monitored. Although many of the differences between ^{1}H-NMR spectra obtained from control and toxin-treated rats can be readily observed without detailed mathematical analysis [19], perturbation in the levels of metabolites present at low concentration may be equally diagnostic and important in understanding the biochemical sequelae following a toxic insult. For example, several bile acids are excreted in elevated amounts in the urine of rats treated with the hepatotoxin, galactosamine [17]. Although these bile acids are present in relatively low concentrations, the pattern of bile acids is indicative of the site of toxicity within the liver and have more diagnostic potential than other more dominant spectral changes, such as the marked reduction in urinary citrate [17]. The use of mathematical models allows an unbiased assessment of the response of metabolites, regardless of the extent of their contribution to the overall composition of a biofluid.

In the last decade, a major contribution to the NMR spectroscopic analysis of biological samples has been the introduction of high-resolution magic angle spinning (MAS) NMR spectroscopy to the analysis of intact biological tissues. Although ^{1}H-NMR-detected perturbation in biofluids, such as urine and plasma, can give rise to surrogate markers of tissue-specific toxicity and can lend insight into the mechanism of toxicity, it cannot give unequivocal evidence for damage to specific tissues per se. Urine data contains components from metabolic processes throughout the body, and therefore, it is often necessary to analyze tissues directly in order to provide a direct link between the histopathology of a lesion and biofluid NMR spectroscopic data. *In vivo* NMR spectroscopy has been used to investigate abnormal tissue biochemistry, but spectral quality is always severely compromised by the high heterogeneity in the sample causing magnetic field inhomogeneity and the constrained molecular motions of molecules in some tissue compartments, leading to poor resolution. Therefore, NMR spectral analysis of tissues has largely relied upon tissue extraction methods [20]. However, extraction processes result in the loss of tissue components such as proteins and lipids. By spinning solid or semi-solid samples, such as biological tissues, at the magic angle (54.7° relative to the applied magnetic field), several important line-broadening effects are reduced and it is possible to obtain very high quality NMR spectra of whole tissue samples with no sample pre-treatment. At this angle, line-broadening effects due to sample heterogeneity and inherent magnetic field inhomogeneity, residual dipolar couplings and chemical shift anisotropy are reduced by scaling the FID by $(3\cos^2\theta - 1)/2$. High-resolution MAS NMR spectroscopy has been used to characterize the low molecular weight composition of a range of biochemical tissues and organelles, including liver, kidney, brain, heart, adipose and mitochondria and to evaluate the biochemical consequences of several toxins and disease processes [21,22,23•,24,25]. In addition to 'bridging the gap' between hisopathology and biofluid analysis, MAS spectroscopy can be used to visualize dynamic processes and to gain insight into the compartmentalization of metabolites within cellular environments [26].

Application of chemometric analysis to NMR data

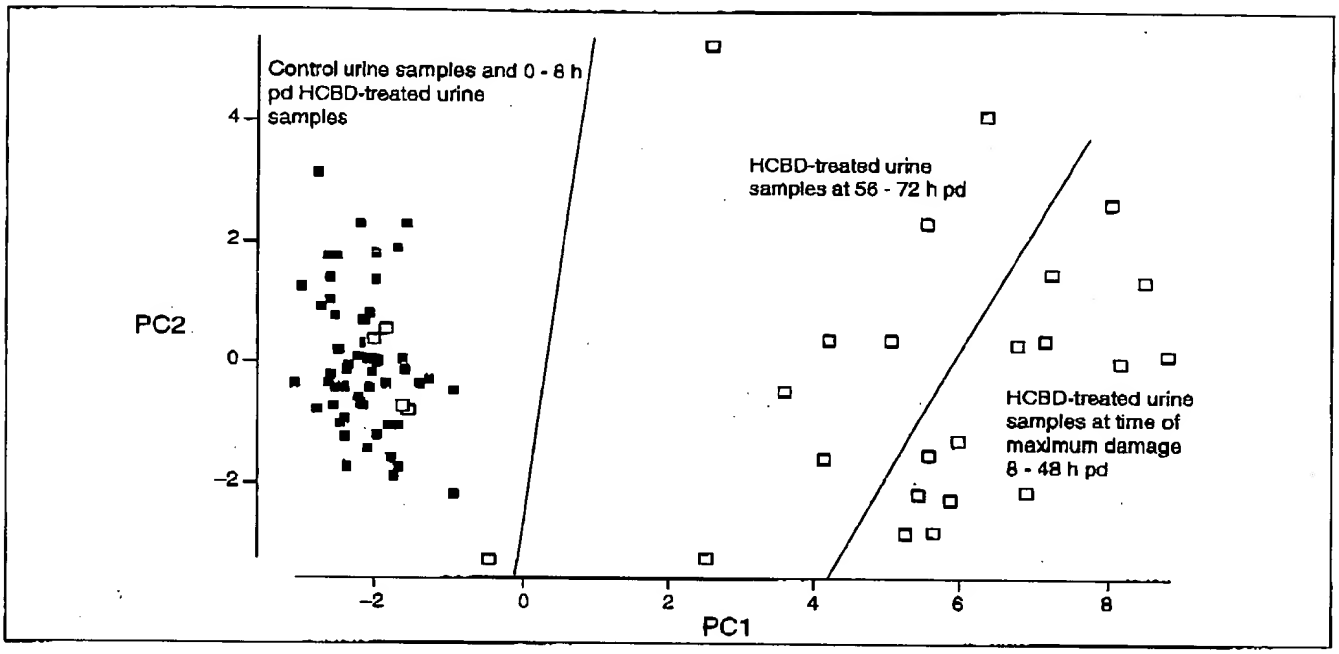
PR and related multivariate statistical approaches can be used to discern significant patterns in complex data sets [27], and are particularly appropriate in situations where there are more variables than samples in the data set, such as is the case with spectral data. The general aim of PR is to classify objects (in this case, ^{1}H-NMR spectra of biofluid or tissue samples) or to predict the origin of objects based on identification of inherent patterns in a set of indirect measurements. PR can be used to reduce the dimensionality of complex data sets via 2- or 3-D mapping procedures, thereby facilitating the visualization of inherent patterns in the data set. Alternatively, multiparametric spectral data can be modeled using PR techniques, so that the class of a sample from an independent data set can be predicted based on a series of mathematical models derived from the original data or 'training set'. Both the theory and the application of the basic mathematical models used in PR have been well-documented [28••,29-31].

Early multivariate approaches to the analysis of ^{1}H-NMR spectra of rat urine, following a chemically-induced toxic insult, involved the use scored or quantitated measurements of selected metabolite signals to indicate an elevation or depletion in the levels of selected urinary metabolites after toxic insult [16•,18,32], followed by appropriate mathematical analysis. However, despite the rudimentary nature of scoring systems, clear relationships between metabolic composition and the dominant site of toxicity could be established and consequent classification of toxins made. These early studies showed that classification of toxins that targeted the renal cortex, renal medulla, liver and testes could be achieved [32]. However, selection and quantification of metabolites is a time-consuming process and involves the *a priori* selection of metabolites, thereby limiting the sensitivity of the NMR-PR approach and imposing an unnecessary degree of subjectivity in metabolite selection. More recent methods of selecting spectral descriptors involve automated approaches that incorporate the whole NMR spectrum, either using computer points or integrated spectral regions. Spectral descriptors can be scaled by a variety of methods in order to optimize data recovery [5••,19]. In combination with PR techniques, ^{1}H-NMR spectroscopy has been used to identify changes in biofluid metabolite concentrations, reflecting site and mechanism-specific toxicity [1•,6,9], to define novel indices of toxic insult [11,12], to evaluate control data [33] and to track progression and regression of toxin-induced lesions over a time period [16•,17]. Furthermore, this metabonomic approach has been shown to be sensitive enough to characterize biochemical differences in urine composition in closely related strains of rat (Han Wistar and Sprague-Dawley) [5•], and therefore, has potential in the evaluation of genetically-modified animals.

One of the most useful and easily applied PR techniques is Principal Components Analysis (PCA), which is a technique that requires no *a priori* knowledge as to the class of the samples. Principal components (PCs) are linear combinations of the original variables and are calculated such that: (i) each PC is orthogonal (uncorrelated) with all other PCs; and (ii) the first PC contains the largest part of the variance of the data set (information content) with subsequent PCs containing correspondingly smaller amounts of variance. Thus a plot of the first two or three PCs gives the 'best' representation, in terms of biochemical variation in the data set in two or three dimensions. An example of applying PCA to the analysis of ^{1}H-NMR urine spectra obtained from control rats and rats treated with a single dose of the nephrotoxin HCBD is given in Figure 3. Three groups of urine samples can be seen in the PC map of ^{1}H-NMR spectra corresponding to those obtained from control rats and rats treated with HCBD prior to the onset of toxicity (0 to 8 h post-dose), those samples obtained over the period of maximum damage (8 to 48 h post-dose), and samples from later time periods (56 to 72 h post-dose) during the onset of recovery.

However, unsupervised chemometric methods, such as PCA, have limited capabilities of classification, particularly where large numbers of classes exist within a data set. Therefore, once evidence of clustering behavior (relating to type or mechanism of toxicity) has been established, supervised methods of analysis can be used to maximize the separation

Figure 3. PC map derived from ^{1}H-NMR spectra of urine obtained over a 7-day time period following the administration of hexachlorobutadiene (200 mg/kg) to a SD rat.



between two or more sample classes and to define features (ie, biochemical markers) that distinguish each class of toxin-treated urine sample from control. These supervised methods include Soft Independent Modeling of Class Analogy (SIMCA), K nearest neighbor (KNN) and neural network analysis [5•].

Development of NMR-PR-based 'expert systems' for toxicological screening

Metabonomic expert systems for the prediction of toxicity can be constructed from a series of mathematical models derived from a training database where the class of toxicity for all samples in the database is known. These multivariate models can be derived from one or more of a range of multivariate statistical methods including PCA, neural network analysis, SIMCA, rule induction and partial least squares (PLS) analysis [5•]. The statistical models are then validated with an independent or 'test' set of samples, where the outcome of toxicity is known, but not used in the mathematical algorithm. Having checked the robustness of the models using a test set, the system can then be used to assess and predict the toxicity of novel xenobiotics. Expert systems can operate at three separate levels:

Level 1. Classification of a sample or organism as 'normal or abnormal'. Classification as abnormal indicates a deviation from the control population and can be caused by numerous factors including toxicity, disease, dietary differences, genetic modification and contamination. This selection of abnormal samples can be achieved automatically 'on-line' and any sample defined as abnormal will undergo further NMR measurements or multivariate statistical analysis with a view to ascertaining the nature of the abnormality.

Level 2. Classification of toxicity. Samples identified as being dissimilar to matched control samples can be fitted to a series of mathematical models that define the multivariate boundaries for known classes of toxicity. Therefore, biofluid or tissue samples from experimental animals treated with novel drugs can be tested to ascertain if the drug induces biochemical effects that would infer a particular site or mechanism of toxicity.

Level 3. Identification of the biomarkers. The metabolites that differ between biofluid samples obtained from drug-treated and control rats can be elucidated giving an insight into possible mechanisms of toxicity or dysfunction. NMR-PR-based expert systems should provide a practical toxicological probe with which to evaluate the potential of novel pharmaceutical compounds.

Conclusions

NMR-based metabonomics can be used to address a large range of toxicological, clinical and environmental problems. Current technology enables the generation of substantial amounts of metabolic data from even simple ^{1}H-NMR experiments on whole biofluids, giving a comprehensive representation of the biochemical processes occurring in whole organisms under different physiological and pathophysiological conditions. Metabonomics has already become a recognized part of toxicological assessment in the pharmaceutical industry. Ongoing developments in instrumentation, multivariate statistical techniques and the interfacing of 'user-friendly' software, should serve to make metabonomics an integral component of toxicological screening and lead compound selection in the pharmaceutical industry.

References

•• of outstanding interest
• of special interest

1. Nicholson JK, Wilson ID: **High resolution proton NMR spectroscopy of biological fluids.** *Prog NMR Spectrosc* (1989) 21:444-501.

• *A comprehensive review of the applications of NMR spectroscopic analysis of biofluids in the fields of toxicology and drug metabolism.*

2. Spraul M, Hofmann M, Ackermann M, Nicholls AW, Damment SJ, Haselden JN, Shockcor JP, Nicholson JK, Lindon JC: **Flow injection ^{1}H NMR spectroscopy combined with pattern recognition methods: implications for rapid structural studies and high throughput biochemical screening.** *Anal Comm* (1997) 34:339-341.

3. Sinclair B: **Everything's great when it sits on a chip: A bright future for DNA arrays.** *Scientist* (1999) 13(11):18-20.

4. Gelsow MJ: **Proteomics: One small step for a digital computer, one giant leap for humankind.** *Nature Biotechnol* (1998) 16(2):206-208.

5. Nicholson, JK, Lindon, JC, Holmes, E: **'Metabonomics': Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data.** *Xenobiotica* (1999) 11:1181-1189.

•• *This article is the first to define the origin of NMR-based metabonomics and discusses the philosophy behind the discipline.*

6. Gartland KP, Bonner FW, Nicholson JK: **Investigations into the biochemical effects of region-specific nephrotoxins.** *Mol Pharmacol* (1989) 35:242-250.

7. Bales JR, Higham DP, Howe I, Nicholson JK, Sadler PJ: **Use of high-resolution proton nuclear magnetic resonance spectroscopy for rapid multi-component analysis of urine.** *Clin Chem* (1984) 30:426-432.

8. Nicholson JK, Buckingham MJ, Sadler PJ: **High resolution ^{1}H NMR studies of vertebrate blood and plasma.** *Biochem J* (1983) 211:605-615.

9. Nicholson JK, Timbrell JA, Sadler PJ: **Proton NMR spectra of urine as indicators of renal damage: Mercury-induced nephrotoxicity in rats.** *Mol Pharmacol* (1985) 27:644-651.

10. Nicholson JK, O'Flynn MP, Sadler PJ, Macleod AF, Juul SM, Sönksen PH: **Proton-nuclear-magnetic-resonance studies of serum, plasma and urine from fasting normal and diabetic subjects.** *Biochem J* (1984) 217:365-375.

11. Nicholson JK, Higham DP, Timbrell JA, Sadler PJ: **Quantitative ^{1}H NMR urinalysis studies on the biochemical effects of cadmium in the rat.** *Mol Pharmacol* (1989) 36:398-404.

12. Sanins SM, Timbrell JA, Elcombe C, Nicholson JK: **Hepatotoxin-induced hypertaurinuria: a proton NMR study.** *Arch Toxicol* (1990) 64:407-411.

13. Connor SC, Hughes MG, Moore G, Lister CA, Smith SA: **Antidiabetic efficacy of BRL 49653 a potent orally active insulin sensitizing agent, assessed in the C57BL/KsJ *db/db* diabetic mouse by non-invasive ^{1}H NMR studies in urine.** *J Pharm Pharmacol* (1997) 49:336-344.

14. Howells SL, Maxwell RJ, Peet AC, Griffiths JR: **An investigation of tumor ^{1}H nuclear magnetic resonance spectra by the application of chemometric techniques.** *Mag Reson Med* (1992) 28:214-236.

15. Waterfield CJ, Turton JA, Scales MD, Timbrell JA: **Investigations into the effects of various hepatotoxic compounds on urinary and liver taurine levels in rats.** *Arch Toxicol* (1993) 67:244-254.

• *The role of taurine as a biomarker of liver toxicity is assessed in this paper.*

16. Holmes E, Bonner FW, Sweatman BC, Lindon JC, Beddell CR, Rahr E, Nicholson JK: **Nuclear magnetic resonance spectroscopy and pattern recognition analysis of the biochemical processes associated with the progression of and recovery from nephrotoxic lesions in the rat induced by mercury(II) chloride and 2-bromoethanamine.** *Mol Pharmacol* (1992) 42:922-930.

• *This article focuses on the use of pattern recognition to follow the development of acute toxic lesions in the rat over a 7-day time period.*

17. Beckwith-Hall BM, Nicholson JK, Nicholls AW, Foxall PJ, Lindon JC, Connor SC, Abdi M, Connelly J, Holmes E: **Nuclear magnetic resonance spectroscopic and principal components analysis investigations into biochemical effects of three model hepatotoxins.** *Chem Res Toxicol* (1998) 11:260-272.

18. Gartland KP, Beddell CR, Lindon JC, Nicholson JK: **Application of pattern recognition methods to the analysis and classification of toxicological data derived from proton nuclear magnetic resonance spectroscopy of urine.** *Mol Pharmacol* (1991) 39:629-642.

19. Holmes E, Nicholls AW, Lindon JC, Ramos S, Spraul M, Neidig P, Connor SC, Connelly J, Damment SJ, Haselden JN, Nicholson JK: **Development of a model for classification of toxin-induced lesions using ^{1}H NMR spectroscopy of urine combined with pattern recognition.** *NMR Biomed* (1998) 11:1-10.

20. Maxwell RJ, Martinez-Perez I, Cerdan S, Cabanas ME, Arus C, Moreno A, Capdevila A, Ferrer E, Bartomeus F, Aparicio A, Conesa G, Roda JM, Carcellar F, Pascual JM, Howells SL, Mazucco R, Griffiths JR: **Pattern recognition analysis of ^{1}H NMR spectra from perchloric acid extracts of human brain tumor biopsies.** *Magn Reson Med* (1998) 39:869-877.

21. Garrod S, Humpfer E, Spraul M, Connor SC, Polley S, Connelly J, Lindon JC, Nicholson JK, Holmes E: **High-resolution magic angle spinning ^{1}H NMR spectroscopic studies on intact rat renal cortex and medulla.** *Magn Reson Med* (1999) 41:1108-1118.

22. Moka D, Schicha H, Vorreuther R, Spraul M, Foxall PJ, Nicholson JK, Lindon JC: **Magic angle spinning proton nuclear magnetic resonance spectroscopic analysis of intact kidney tissue samples.** *Anal Comm* (1997) 34:107-110.

23. Tomlins A, Foxall PJ, Lindon JC, Lynch MJ, Spraul M, Everett J, Nicholson JK: **High resolution magic angle spinning ^{1}H nuclear magnetic resonance analysis of intact prostatic hyperplastic and tumor tissues.** *Anal Comm* (1998) 35:113-115.

• *Recent developments in high-resolution magic angle spinning probes have revolutionized the ^{1}H-NMR spectroscopic analysis of intact tissues. Using this technique it is now possible to obtain spectra of biological tissues which have almost the same resolution as those obtained from tissue extracts, without losing any biochemical components in the process of tissue extraction.*

24. Cheng LL, Lean CL, Bogdanova A, Wright SC Jr, Ackerman JL, Brady TJ, Garrido L: **Enhanced resolution of proton NMR spectra of malignant lymph nodes using magic-angle spinning.** *Magn Reson Med* (1996) 36:653-658.

25. Moka D, Vorreuther R, Shicha H, Humpfer E, Lipinski M, Spraul M, Foxall PJ, Nicholson JK, Lindon JC: **Biochemical classification of kidney carcinoma biopsy samples using magic-angle-spinning ^{1}H nuclear magnetic resonance spectroscopy.** *J Pharm Biomed Anal* (1998) 17:125-132.

26. Humpfer E, Spraul M, Nicholls AW, Nicholson JK, Lindon JC: **Direct observation of resolved intracellular and extracellular water signals in intact human red blood cells using ^{1}H MAS NMR spectroscopy.** *Magn Reson Med* (1997) 38:334-336.

27. Manley BF (Ed): *Multivariate Statistical Methods: A Primer.* Chapman and Hall, London (1986).

28. El-Deredy W: **Pattern recognition approaches in biomedical and clinical magnetic resonance spectroscopy: A review.** *NMR Biomed* (1997) 10:99-124.

•• *A wide range of the applications of NMR-PR analysis, and the potential role of this technique in the biomedical and pharmaceutical sciences are covered in this review.*

29. Beebe KR, Pell RJ, Seahsholt MB (Eds): *Chemometrics: A Practical Guide.* John Wiley & Sons, New York (1998).

30. Jurs PC: **Pattern recognition used to investigate multivariate data in analytical chemistry.** *Science* (1986) 232:1219-1224.

31. Erikseon L, Johansson E, Kettanah-Wold N, Wold S: **Introduction to multi and megavariate data analysis using projection methods (PCA and PLS).** Umetrics AB, Mälmo, Sweden (1999). *http://www.umetrics.com*

32. Anthony ML, Sweatman BC, Beddell CR, Lindon JC, Nicholson JK: **Pattern recognition classification of the site of nephrotoxicity based on metabolic data derived from proton nuclear magnetic resonance spectra of urine.** *Mol Pharmacol* (1994) 46:199-211.

33. Holmes E, Foxall PJ, Nicholson JK, Neild GH, Brown SM, Beddell C, Sweatman BC, Rahr E, Lindon JC, Spraul M, Neidig P: **Automatic data reduction and pattern recognition methods for analysis of ^{1}H nuclear magnetic resonance spectra of human urine from normal and pathological states.** *Anal Biochem* (1994) 220:284-296.

458208

Jacob, Rebecca (ASRC)

From: STIC-ILL
Sent: Tuesday, August 05, 2003 3:06 PM
To: Jacob, Rebecca (ASRC)
Subject: FW: 09890973

-----Original Message-----
From: Mellerson, Kendra
Sent: Tuesday, August 05, 2003 3:06 PM
To: STIC-ILL
Subject: FW: 09890973

-----Original Message-----
From: Gakh, Yelena
Sent: Tuesday, August 05, 2003 2:33 PM
To: STIC-EIC1700
Subject: 09890973

Dear Kendra:

please order one more list:

10. TITLE: "The identification of novel biomarkers of renal toxicity using automatic data reduction techniques and PCA of proton NMR spectra of urine"
AUTHOR(S): *Holmes, Elaine; Nicholson, Jeremy K.; Nicholls, Andrew W.; Lindon, John C.; Connor, Susan C.; Polley, Stephen; Connelly, John*
CORPORATE SOURCE: Birkbeck College, Department of Chemistry, University of London, London, SW7 2AZ, UK
SOURCE: **Chemometrics and Intelligent Laboratory Systems (1998), 44(1,2), 245-255**

Thank you,

Yelena

Yelena G. Gakh, Ph.D.

Patent Examiner
USPTO, cp3/7B-08
(703)306-5906

ELSEVIER

Chemometrics and Intelligent Laboratory Systems 44 (1998) 245–255

Chemometrics and intelligent laboratory systems

The identification of novel biomarkers of renal toxicity using automatic data reduction techniques and PCA of proton NMR spectra of urine

Elaine Holmes [a,*], Jeremy K. Nicholson [a], Andrew W. Nicholls [a], John C. Lindon [a], Susan C. Connor [b], Stephen Polley [c], John Connelly [b]

[a] *Section of Biological Chemistry, Biomedical Sciences, Imperial College of Science, Technology and Medicine, University of London, Alexander Flemming Building, Exhibition Road, London SW7 2AZ, UK*

[b] *Department of Analytical Sciences, SmithKline Beecham Pharmaceuticals, New Frontiers Science Park, Essex, UK*

[c] *Department of Safety Assessment, SmithKline Beecham Pharmaceuticals, New Frontiers Science Park, Essex, UK*

Received 19 August 1997; revised 20 May 1998; accepted 22 June 1998

Abstract

Early detection of drug-induced toxic lesions is of considerable importance in the pharmaceutical industry. Many drugs and toxins produce characteristic patterns of biochemical perturbations in the urinary profile related to the site or mechanism of the lesion. ^{1}H nuclear magnetic resonance (NMR) spectroscopy of biofluids has been shown to be a useful technique for characterising such lesions. We present here an efficient approach to the analysis and classification of complex urine NMR spectra obtained from rats treated with various nephrotoxins (glomerular, papillary and proximal tubular) based on the automatic generation of descriptors for the spectra with subsequent PCA. Urinalysis was performed using 600 MHz ^{1}H NMR spectroscopy and the site of renal lesion was confirmed by renal histology. A plot of the first three PCs showed distinct clustering of urine samples reflecting the site of toxicity within the kidney. Interrogation of the eigenvectors showed which NMR spectral regions contributed most to the separation of classes. These regions were examined visually for perturbations in metabolite profile and sets of 'marker' metabolites that characterised tissue-specific lesions were defined. These studies have shown that automatic data reduction of the spectra followed by multivariate techniques such as principal components analysis (PCA) is a reliable method for screening for biomarkers of organ or tissue-specific chemically-induced lesions.

Keywords: ^{1}H NMR spectroscopy; Nephrotoxin; Principal components analysis

Abbreviations: 2-Bromoethanamine hydrobromide (BEA); Hexachlorobutadiene (HCBD); Lead acetate (PbAc); Mercury II chloride ($HgCl_2$); Nuclear magnetic resonance (NMR) spectroscopy; Principal components analysis (PCA); Propyleneimine (PI); Puromycin aminonucleoside (PAN); Sodium chromate ($NaCrO_4$); 1,1,2-Trichloro-3,3,3-trifluoro-1-propene (TCTFP)

* Corresponding author. Tel.: +44 410 259 113.

0169-7439/98/$ - see front matter
PII: S0169-7439(98)00110-5

1. Introduction

Early identification of drug toxicity is an important factor in facilitating the selection of lead compounds for drug development. The interaction of toxins with cells or tissues can cause perturbation of the ratios and concentrations of endogenous biochemicals involved in key metabolic pathways [1]. In order to maintain homeostasis, and to adjust for the changes in tissue biochemistry, the compositions of the body fluids are altered accordingly. High field ^{1}H NMR spectroscopy has been shown to be a useful method of monitoring perturbed biofluid profiles since a large range of low molecular weight metabolites can be viewed simultaneously [1,2]. Although some xenobiotics induce widespread organ toxicity, many others have been shown to highly specific in targeting specific tissues [1,3–5]. Previous studies have shown that there is a relationship between site of toxic lesion and the pattern of metabolic perturbations in the ^{1}H NMR urine profiles [2,6]. For example, increased urinary excretion of glucose, amino acids and organic acids have been shown to indicate damage to the renal proximal tubule in the S_3 region [2].

Biofluid ^{1}H NMR spectra are inherently complex, typically each spectrum being comprised of ca. 64–128 k data points showing thousands of partially overlapped resonances when Fourier transformed into the frequency domain. The complexity of biofluid spectra obtained at high frequencies (500 MHz or greater) can lead to difficulties in data interpretation. To provide an aid to spectral interpretation, various statistical methods of handling biofluid data and accessing latent spectral information have been investigated.

Previous multivariate analysis of NMR urine spectra, obtained from rats following a chemically-induced toxicological insult, involved the use of techniques such as hierarchical cluster analysis and principal components analysis (PCA) of scored or quantitated measurements of selected metabolite signals [7–9]. Even in these early studies a clear relationship between metabolic composition and the dominant site of toxicity was established. Moreover, a distinction could be made between the biochemical effects of toxins which predominantly targeted the renal cortex, renal medulla, liver and testes [7–9]. However, selection and quantitation of metabolites is a time consuming process and involves prior assumptions as to the comparative importance of endogenous metabolites in indicating toxic effect. The a priori selection of such metabolites can result in altered concentrations of other low level species being overlooked. These changes in the levels of low concentration metabolites may be potentially more important in terms of indicating a site or mechanism of toxicity. Extension of these initial studies led to the adoption of automated data reduction procedures whereby the NMR spectrum was divided into regions of equal chemical shift ranges and the integrals within those ranges calculated. This technique was used to characterise human urine samples obtained from patients with inborn errors of metabolism and to establish the range of normal physiological variance in human urine [10,11]. A similar automatic data reduction procedure has been used to classify NMR spectral data obtained from various types of tumour tissue extracts where the peak height was calculated at fixed intervals prior to hierarchical cluster analysis and PCA [12]. Other examples of the successful application of chemometric analysis to automatically-reduced NMR spectral data can also be found in the food industry where NMR-PR models have been used to classify products such as German wines and apple juice according to the region of origin [13,14].

The aim of the current study was to investigate the potential of automatic data reduction and PCA in classifying the site and severity of chemically induced renal lesions. To this end 15 groups of rats (n = 5 per group) each received a single dose of a nephrotoxin whose site of action within the nephron was known (Table 1). In addition, three groups of control animals (n = 5 per group) each received the dosing vehicle only. The development of a lesion is a time-dependent process and the various phases of lesion development and recovery may be associated with different metabolic profiles [9]. Therefore, in order to achieve a more complete biochemical profile for each nephrotoxin, urine samples were collected at various time points for up to 7 days after treatment. All urine samples were analysed using ^{1}H NMR spectroscopy and automated data reduction procedures were then employed to reduce each spectrum into a series of discrete integrated regions. PCA was used to map the samples (based on the integrated regions), and to ascertain whether this tech-

Table 1
Nephrotoxic compounds administered together with respective dose levels and areas of effect

Compound	Region of effect	Dose (mg/kg)	Severity of lesion (as confirmed by histopathology)
$HgCl_2$	Cortex—proximal tubule S_3	0.75	*****
HCBD	Cortex—proximal tubule S_3	200	*****
Uranyl nitrate	Cortex—proximal tubule S_3	10	****
TCTFP	Cortex—proximal tubule S_3	20	*
Cisplatin	Cortex—proximal tubule S_3/distal tubule	6	***
Cephaloridine	Cortex—proximal tubule $S_{1/2}$	750	**
$NaCrO_4$	Cortex—S_1	20	**
Sodium fluoride	Cortex—proximal tubule	35	***
BEA	Papilla	150	****
PI	Papilla	0.016	**
Adriamycin	Glomerulus	5	*
PAN	Glomerulus and proximal tubules	150	***glomerulus *****tubules
Amphotericin B	Distal and proximal tubules	10	*
PbAc	Liver/kidney	98	**
$CdCl_2$	Testicular/kidney	1	**kidney ****testicles

*No histological changes observed.
**Mild necrosis.
***Mild/moderate necrosis.
****Moderate/severe necrosis.
*****Severe necrosis (involving up to 80% of the targeted tissue).

nique would be suitable for relating changes in specific spectral regions to the topographical area of effect within the kidney.

2. Experimental

2.1. Treatments and sample preparation

A single i.p. dose of either control vehicle (0.9% saline or corn oil) or one of 15 nephrotoxic compounds was administered to male Sprague–Dawley (SD) rats ($n = 5$ per group). The nephrotoxins administered and their respective dose levels, together with the site of action within the nephron, are given in Table 1. Each animal was housed individually in a metabolism cage and urine samples were collected at various time intervals (pre-dose and 0–8 h, 8–24 h, 24–32 h, 32–48 h, 48–72 h, 72–96 h, 96–120 h, 120–144 h and 144–168 h after treatment). Urine samples were centrifuged at 3000 rpm for 10 min in order to remove particulate contaminants and the samples were stored at −40°C pending NMR spectroscopic analysis.

In order to minimise variations in the pH of the urine samples, 200 μl of a buffer solution (0.2 M Na_2HPO_4/0.2 M NaH_2PO_4, pH = 7.4) was mixed with 400 μl of urine in a micro-container. The resulting solution was left to stand for 10 min and then centrifuged at 13,000 rpm for 10 min to remove any precipitate. A total of 500 μl of the supernatant was placed into a 5 mm o.d. NMR tube (Wilmad 507PP). A field-frequency lock was provided by adding 100 μl ^2H_2O solution to the sample in the NMR tube.

2.2. ^1H NMR spectroscopy of urine

^1H NMR spectra were measured at 600.13 MHz on a Bruker DRX-600 spectrometer. The water resonance was suppressed using the first increment of a NOESY pulse sequence with irradiation during a 3's relaxation delay and also during the 100 ms mixing time. Typically 64 free induction decays (FIDs) were collected into 64k data points using a spectral width

of 7002.8 Hz, an acquisition time of 4.68 s and a total pulse recycle time of 7.68 s. Prior to Fourier transformation (FT) the FIDs were zero-filled to 128k and an exponential line broadening factor of 0.3 Hz was applied. All spectra were phase-corrected and referenced to the CH_3 resonance of creatinine at δ 3.05. A baseline correction factor was also applied to each spectrum using a simple polynomial curve fit.

2.3. NMR data reduction procedures and pattern recognition analysis

Each NMR spectrum was segmented into 250 chemical shift regions of 0.04 ppm width using the software package AMIX (Analysis of Mixtures, version 2.0, Bruker Analytische Messtechnik, Rheinstetten, Germany). The integral was calculated for each

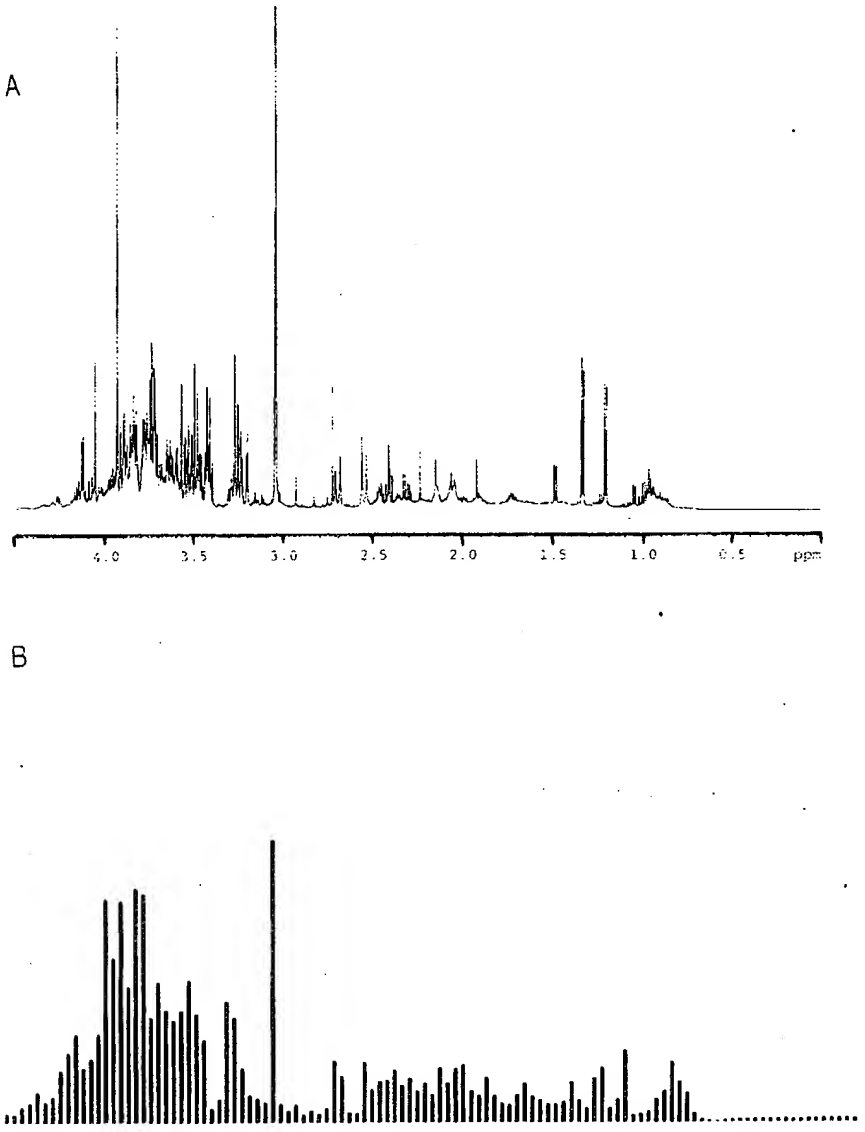


Fig. 1. 600 MHz ^{1}H NMR spectrum of urine obtained from a rat treated with hexachloro-1,3-butadiene (A) and the corresponding data-reduced spectrum (B).

of the spectral regions (a typical result is shown in Fig. 1) and these data were imported into SAS version 6.11 (SAS Institute, Cary, NC, USA). A segment width of 0.04 ppm was chosen in order to accommodate the effects of minor variations in pH which would lead to small changes in chemical shift of certain metabolites. Any background offset in each NMR spectrum was corrected by subtraction of the mean integral for the first 20 regions since these regions were known to contain no NMR resonances. Regions where the integral value amounted to less than five times the calculated noise level were also discarded. Those integrated regions which contained resonances from either the residual water or urea were removed from the data table in order to eliminate both the variation in water suppression and the variation in the integral of the urea signal due to partial cross saturation via the solvent-exchanging protons. It was also necessary to remove all spectral areas that contained resonances arising from the xenobiotic-compounds administered or their metabolites in order to allow classification of toxicity solely based on endogenous markers. Where drug metabolites dominated a substantial proportion of the spectrum for a transitory period, spectra collected over the period in question were discarded.

The remaining integral values were scaled to the total of the summed integrals for each spectrum with a view to compensating in part for the differences in concentration (osmolarity) between individual urine samples.

Initially, a PCA was performed separately for each toxin and the relevant control. PCA plots of the first three components allowed visualisation of the data and to establish whether there were any intrinsic toxin-related differences in the metabolic composition of the urine. The PC loadings were examined in order to determine which variables contributed most to the PC in which separation was observed and hence to indicate which spectral regions were most dominant in separating classes. The ^{1}H NMR spectra were subsequently examined with a view to identifying possible markers of toxicity type within these regions.

Mean values of the NMR descriptors were calculated for each class of urine samples at each time point and plots of PC1 vs. PC2 for the mean data were constructed. These maps gave an indication of the progression of the lesion through time and were used to identify the time points of maximum biochemical effect for each toxin. Data corresponding to the time points of maximum biochemical effect for each toxin were combined and a full PCA performed. PCA maps of the first three PCs were produced and examined for inherent clustering behaviour that could be related to the site or mechanism of toxicity.

3. Results and discussion

The ^{1}H NMR derived metabolic profiles appeared to be unique for each of the 15 toxins studied. However, in many cases, particularly for the S_3 cortical toxins, the urine profiles were also found to be strongly characteristic of the discrete topographical region of the nephron. Typical ^{1}H NMR spectra obtained from animals treated with selected nephrotoxins are illustrated in Fig. 2. The time of onset of biochemical changes in the urine was toxin dependent. Some compounds such as sodium chromate caused an immediate alteration of biochemical profile with increased urinary concentrations of glucose and acetate followed by a return to control levels by 24 h post dose. Other compounds such as puromycin aminonucleoside did not show signs of glomerular toxicity until 96 h after treatment when lipiduria and proteinuria were apparent. These effects were confirmed as toxicity-related by independent histological examination.

PCA proved to be a useful and rapid means of establishing whether the urine spectra obtained from rats treated with a particular nephrotoxin were different from those obtained from control animals and also of identifying at which time points maximum biochemical effect occurred. When mapped individually, all toxins studied were distinguishable from controls at one or more of the time periods. For example, clear separation of samples obtained from uranyl nitrate-treated rats and control rats at all except one (0–8 h post dose) time point is shown in the PC map in Fig. 3A. This would indicate that the onset of toxicity occurred 8 h after treatment and that regeneration of the tissue was not complete by the end of the study (168 h post dose). Histology of the kidney at 168 h post dose confirmed that severe tubular nephropathy was still present at this time.

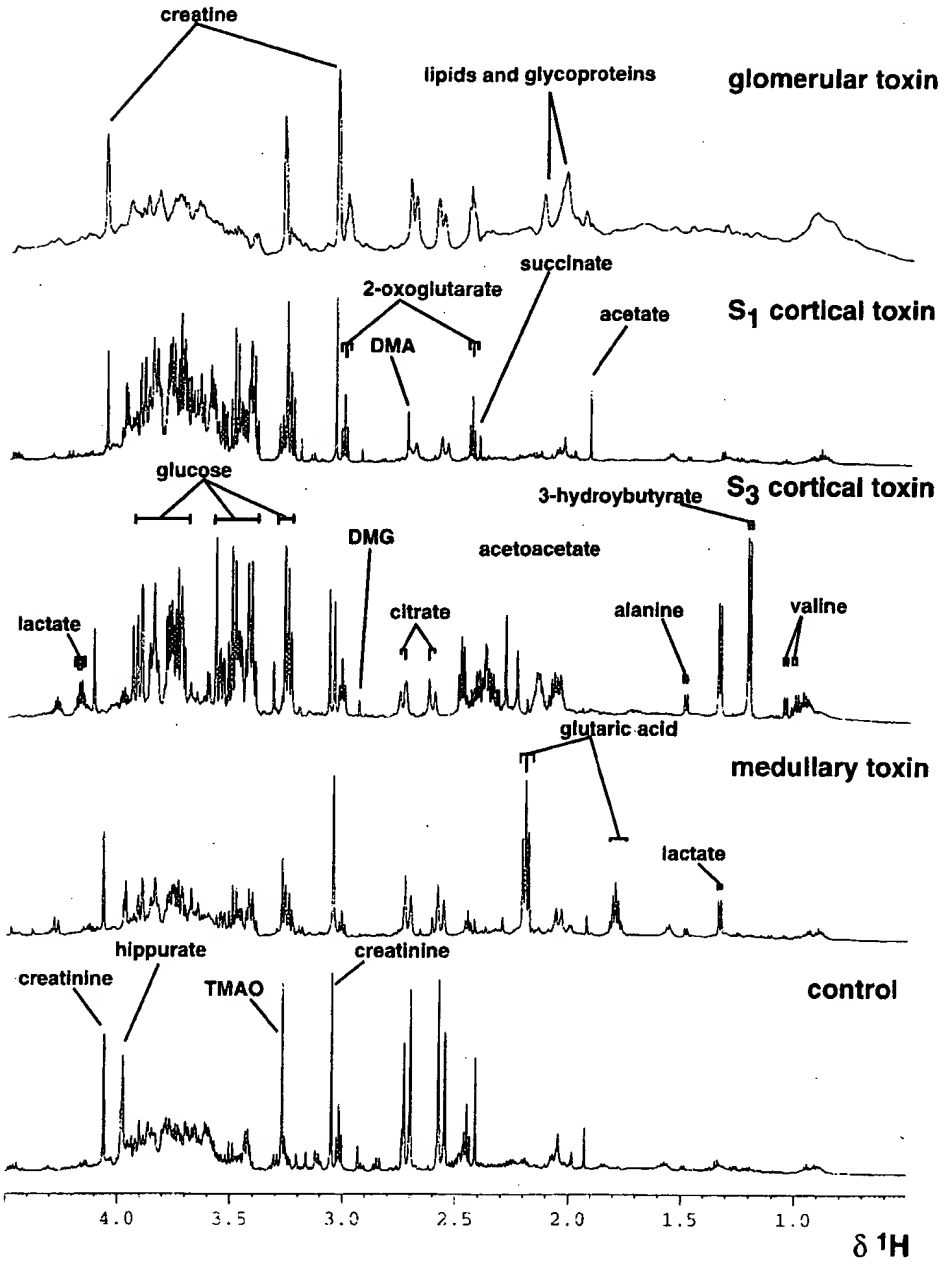


Fig. 2. 600 MHz ^{1}H NMR spectra of urine obtained from a control rat and from rats treated with model compounds that target differe regions of the nephron: bromoethanamine (renal papillary toxin), hexachloro-1,3-butadiene (renal cortical toxin, S_3). sodium chromate (r nal cortical toxin, S_1) and puromycin aminonucleoside (glomerular toxin). Abbreviations: dimethylamine (DMA), dimethylglycine (DMG 2-oxoglutarate (2-OG) and trimethylamine-*N*-oxide (TMAO).

PCA was repeated on group mean data in order to simplify the maps and to establish a biochemical trajectory of effect. Previous work has shown that the onset, progression and recovery from toxin-induce lesions can be efficiently monitored and compared b using PCA to construct a mean trajectory [9]. Th

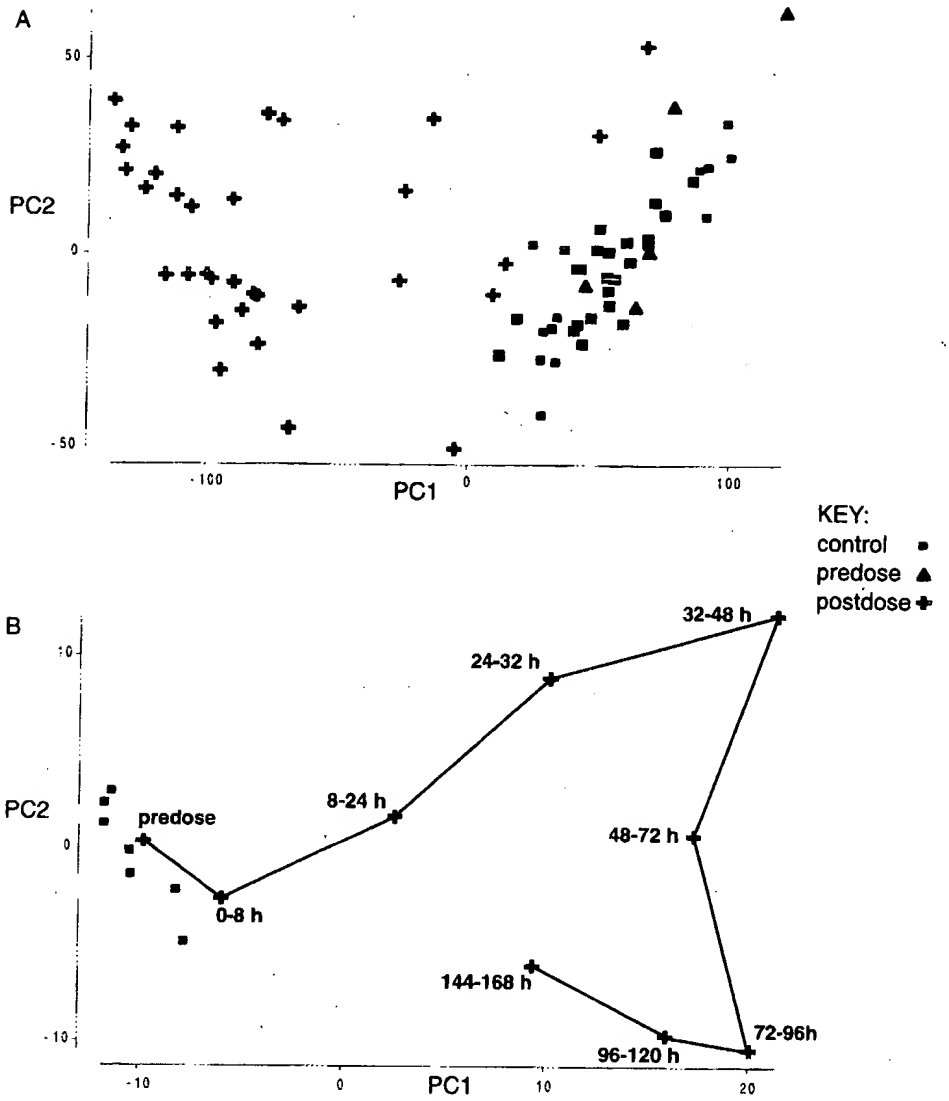


Fig. 3. A plot of PC1 vs. PC2 based on the ^{1}H NMR integral regions for (A) individual urine samples obtained from rats treated with uranyl nitrate and (B) the mean data for the same samples showing a time-related trajectory.

time points of maximum biochemical effect were established by identifying the time points at which the urinary NMR descriptors reached a maximum distance from the pre-dose position in the plot of PC1 vs. PC2. The time course of uranyl nitrate induced toxicological injury is shown in Fig. 3B.

The data set was then constructed to include only those NMR descriptors from urine samples collected at time points associated with maximal biochemical effect. Analysis of the eigenvectors for PC1 and PC2 indicated which spectral regions were predominantly responsible for separation between classes at this time point. The selected spectral regions were examined and potential markers of toxic effect identified. Although no two toxins produced exactly the same pattern of loadings, similarities in patterns were apparent between classes of toxin that affected the same site. For example, regions of the spectrum containing

Table 2
Key endogenous markers of nephrotoxic insult (identified from the PC loadings)

Compound*	Ac	AcAc	ala	cit	cm	crt	DMA	DMG	for	glu	hip	L	lac	MA	3OHB	2-OG	Pr	suc	tau	TMAO	tyr	val
$HgCl_2$	1+	1+	3+	3−	1−	1+	0	1−	1+	3+	3−	0	3+	0	3+	3−	0	3−	0	0	1+	3+
HCBD	0	2+	3+	3−	1−	1+	0	1−	0	3+	3−	0	3+	0	3+	3−	0	3−	0	0	1+	3+
Uranyl nitrate	1+	2+	3+	3−	1−	1+	0	0	0	3+	3−	0	3+	0	3+	3−	0	3−	0	0	1+	3+
TCTFP	0	2+	2+	2−	1−	1+	0	1−	0	2+	1−	0	1+	0	0	2−	0	0	0	0	0	0
Cisplatin	0	2+	2+	3−	1−	3+	0	0	0	2+	3−	0	2+	0	2+	2−	0	0	3+	3−	1+	2+
Adrimycin	0	0	0	3−	0	3+	0	0	0	0	1−	1+	0	0	0	0	1+	3−	2+	3−	0	0
PAN	1+	0	0	3−	0	3+	1+	2−	1+	0	1−	3+	0	0	0	2−	3+	3−	3+	3−	0	0
$NaCrO_4$	3+	0	0	2−	0	1+	0	0	0	3+	2−	0	0	0	0	0	0	0	0	0	0	0
Cephaloridine	0	0	0	2−	0	0	0	0	0	0	2−	0	0	0	0	2−	0	2−	0	2−	0	0
BEA	0	0	1+	1−	0	1+	1+	0	0	1+	1−	0	1+	2+	0	1−	0	1−	0	1−	0	0
PI	0	0	0	1−	0	0	0	0	0	1+	1−	0	0	0	0	1−	0	0	1+	1−	0	0
Amphotericin B	0	0	0	1−	0	0	0	0	0	1+	0	0	1+	0	0	1−	0	0	0	0	0	0
NaF	2+	0	2+	2−	1−	0	0	0	0	2+	1−	0	1+	0	0	2−	0	0	0	0	0	0
PbAc (liver)	3+	1+	0	2−	0	3+	0	1−	1+	0	3−	0	0	0	0	2−	0	2−	2+	2−	0	0
$CdCl_2$ (testicular)	0	0	0	2−	0	2+	0	3−	0	0	0	0	0	0	0	1−	0	0	1+	0	0	0

0: No change in integral value for the region containing the selected metabolite.
1: 20–50% change in integral value for the region containing the selected metabolite.
2: 50–200% change in integral value for the region containing the selected metabolite.
3: > 200% change in integral value for the region containing the selected metabolite.
+: Increase in integral value (compared to control value).
−: Decrease in integral value (compared to control value).

Fig. 4. chemica of outlie

reso buty foun toxir nitra parti ters obtai those

resonances derived from glucose, lactate, 3-hydroxybutyrate, hippurate, citrate and 2-oxoglutarate were found to be among the most significant for all S_3 toxins which caused severe lesions (Table 1; uranyl nitrate, $HgCl_2$, HCBD). For the glomerular toxins, particularly puromycin aminonucleoside, two clusters of samples were observed relating to the samples obtained between 24 and 48 h after treatment and those samples obtained after 72 h post dose. The ^1H NMR spectra showed that there were two sets of spectral markers. At early time points (24–48 h p.d.) elevation in the concentrations of taurine and creatine occurred with a concomitant depletion of citrate, 2-oxoglutarate, succinate, hippurate and glucose. This pattern of metabolite change would suggest that some degree of liver damage had occurred since increased urinary taurine has been associated with hepatotoxicity [15]. Four days after the administration of

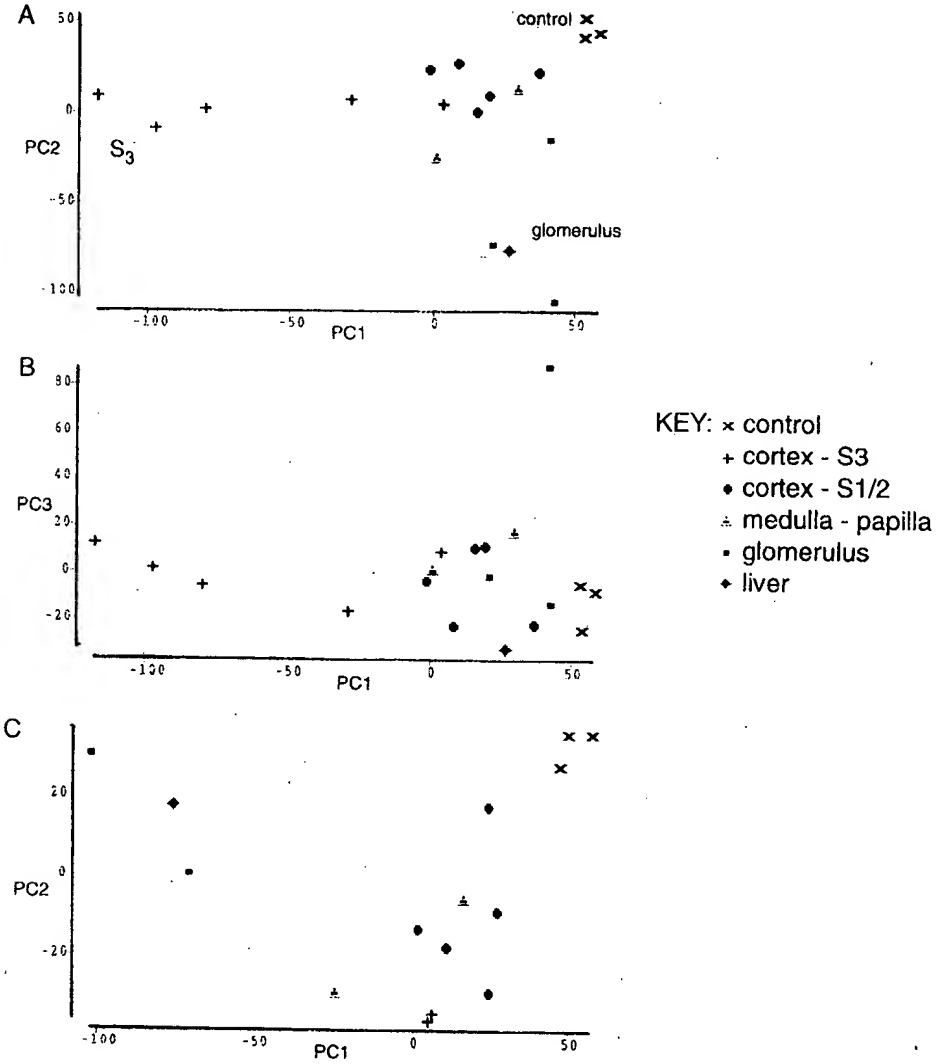


Fig. 4. A plot constructed from the mean spectral descriptors derived from urine samples that were obtained at the time of maximum biochemical effect: (A) PC1 vs. PC2 for all toxins, (B) PC1 vs. PC3 for all toxins and (C) PC1 vs. PC2 for the same data set after the removal of outliers (HCBD, UN, $HgCl_2$ and puromycin aminonucleoside).

puromycin aminonucleoside, further changes in the metabolite profile were observed relating to glomerular effects as confirmed by histology. These changes included increased excretion of lipid and proteins causing general broadening of spectral resonances. The metabolites common to site of toxicity are listed in Table 2.

Data from all toxins studied were combined and PCA was performed. However, only samples collected at time points associated with maximal metabolic perturbation were included in the combined data set. PCA of the combined data showed several site-related clusters. Most distinct was the cluster of S_3 toxins (Fig. 4), although the coordinates representing the biochemical effects of the glomerular toxins and lead acetate formed a separate cluster. For some of the toxins, 1,1,2-trichloro-3,3,3-trifluoro-1-propene, amphotericin B and adriamycin, no histological evidence of renal damage was found. However, the site of lesion for these compounds is well documented [6,16,17]. ^{1}H NMR-detected alterations in the biochemical composition of the urine samples treated with these nephrotoxins would therefore suggest that early biochemical markers of toxic effect can be observed prior to the development of a lesion.

4. Conclusion

The application of automatic data reduction and PCA to the analysis of ^{1}H NMR urine spectra obtained following a nephrotoxic insult has allowed the identification of key regions, and hence markers of region-specific toxicity. This NMR-PCA methodology has the potential for facilitating the process of determining unsuitable candidates for drug development on the grounds of toxicity.

References

[1] J.K. Nicholson, I.D. Wilson, High Resolution NMR spectroscopy of biofluids, Progress in NMR Spectroscopy 21 (1989) 444–501.

[2] K.P.R. Gartland, F.W. Bonner, J.K. Nicholson, Investigations into the biochemical effects of region-specific nephrotoxins, Molecular Pharmacology 35 (1989) 242–250.

[3] J. Halman, J.S.L. Fowler, R.G. Price, Urinary enzymes, proteinuria and renal function tests in the assessment of nephrotoxicity in the rat, in: P.H. Bach, E.A. Lock (Eds.), Renal Heterogeneity and Target Cell Toxicity, Wiley, Chichester, UK, 1985, pp. 295–298.

[4] S.K. Tandon, Organ toxicity of chromium in biological and environmental aspects of chromium, S. Langard (Ed.), Elsevier Biomedical Press, Amsterdam, 1982, pp. 209–220.

[5] C.D. Klassen, in: C.D. Klassen, M.O. Andur, J. Doull (Eds.), Casarett and Doull's Toxicology: The Basic Science of Poisons, 5th edn., McGraw-Hill, 1996.

[6] M.L. Anthony, C.R. Beddell, J.C. Lindon, J.K. Nicholson, Studies on the comparative toxicity of *S*-(1,2-dichlorovinyl)-L-cysteine, *S*-(1,2-dichlorovinyl)-homocysteine and 1,1,2-trichloro-3,3,3-trifluoro-1-propene in the Fischer 344 rat, Archives of Toxicology 69 (1994) 99–110.

[7] K.P.R. Gartland, S.M. Sanins, J.K. Nicholson, B.C. Sweatman, C.R. Beddell, J.C. Lindon, Pattern recognition analysis of high resolution ^{1}H NMR spectra of urine: a nonlinear mapping approach to the classification of toxicological data, NMR in Biomedicine 3 (1990) 166–172.

[8] M.L. Anthony, B.C. Sweatman, C.R. Beddell, J.C. Lindon, J.K. Nicholson, Pattern recognition classification of the site of nephrotoxicity based on metabolic data derived from proton nuclear magnetic resonance spectra of urine, Molecular Pharmacology 46 (1994) 199–211.

[9] E. Holmes, F.W. Bonner, B.C. Sweatman, J.C. Lindon, C.R. Beddell, E. Rahr, J.K. Nicholson, Nuclear magnetic resonance spectroscopy and pattern recognition analysis of the biochemical processes associated with the progression and recovery from nephrotoxic lesions in the rat induced by mercury II chloride and 2-bromoethanamine, Molecular Pharmacology 42 (1992) 922–930.

[10] E. Holmes, P.J.D. Foxall, J.K. Nicholson, G.H. Neild, S.M. Brown, C.R. Beddell, B.C. Sweatman, E. Rahr, J.C. Lindon, M. Spraul, P. Neidig, Automatic data reduction and pattern recognition methods for analysis of ^{1}H magnetic resonance spectra of human urine from normal and pathological states, Analytical Biochemistry 220 (1994) 284–296.

[11] R.D. Farrant, J.C. Lindon, E. Rahr, B.C. Sweatman, An automatic data reduction and transfer method to aid pattern recognition analysis and classification of NMR spectra, Journal of Pharmaceutical and Biomedical Analysis 10 (1992) 141–144.

[12] S.L. Howells, R.J. Maxwell, A.C. Peet, J.R. Griffiths, An investigation of tumour ^{1}H nuclear magnetic resonance spectra by the application of chemometric techniques, Magnetic Resonance in Medicine 28 (1992) 214–236.

[13] T.W.E. Vogels, A.C. Tas, F. vanden Berg, J. van den Greef, A new method for classification of wines based on proton and carbon-13 NMR spectroscopy in combination with pattern recognition techniques, Chemometrics and Intelligent Laboratory Systems 21 (1993) 249–258.

[14] J. Colquhoun, P.S. Belton, E.K. Kemsley, P. Roma, I. Delgadillo, M.J. Dennis, M. Sharman, E. Holmes, J.K. Nicholson, M. Spraul, Applications of NMR spectroscopy and chemometrics to the analysis of carbohydrate mixtures and

fruit juices, Journal of Magnetic Resonance Analysis 2 (1996) 185–186.
[15] C.J. Waterfield, J.A. Turton, M.D.C. Scales, J.A. Timbrell, Investigations into the effects of various hepatoxic compounds on urinary and liver taurine levels in rats, Archives of Toxicology 67 (1993) 244–254.
[16] R.S. Goldstein, Biochemical heterogeneity and site-specific tubular injury, in: J.B. Hook, R.S. Goldstein (Eds.), Toxicology of the Kidney, Raven Press, New York, 1993, pp. 238–239.
[17] T. Bertani, G. Rocchi, G. Sacchi, Adriamycin-induced glomerulosclerosis in the rat, American Journal of Kidney Diseases 7 (1986) 12–19.

Mellerson, Kendra

From: Gakh, Yelena
Sent: Tuesday, August 05, 2003 2:33 PM
To: STIC-EIC1700
Subject: 09890973

Dear Kendra:

please order one more list:

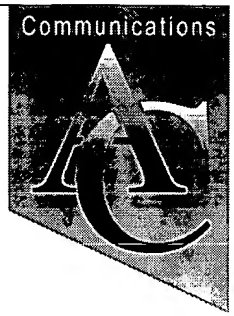
11. TITLE: "Flow injection proton nuclear magnetic resonance spectroscopy combined with pattern recognition methods: implications for rapid structural studies and high throughput biochemical screening"
AUTHOR(S): *Spraul, Manfred; Hofmann, Martin; Ackermann, Michael; Nicholls, Andrew W.; Damment, Stephen, J. P.; Haselden, John N.; Shockcor, John P.; Nicholson, Jeremy K.; Lindon, John C.*
CORPORATE SOURCE: Bruker Analytische Messtechnik GmbH, Rheinstetten, D-76287, Germany
SOURCE: **Analyst (Cambridge, United Kingdom) (1997), 122(11), 339-341**

Thank you,

Yelena

Yelena G. Gakh, Ph.D.

Patent Examiner
USPTO, cp3/7B-08
(703)306-5906



Flow Injection Proton Nuclear Magnetic Resonance Spectroscopy Combined With Pattern Recognition Methods: Implications for Rapid Structural Studies and High Throughput Biochemical Screening

Manfred Spraul[a], Martin Hofmann[a], Michael Ackermann[a], Andrew W. Nicholls[b], Stephen J. P. Damment[c], John N. Haselden[c], John P. Shockcor†[d], Jeremy K. Nicholson[b] and John C. Lindon*[b]

[a] *Bruker Analytische Messtechnik GmbH, Silberstreifen, D-76287 Rheinstetten, Germany*
[b] *Department of Chemistry, Birkbeck College, University of London, Gordon House, 29 Gordon Square, London UK WC1H 0PP. E-mail: jcl@chem.bbk.ac.uk*
[c] *Department of Toxicology, GlaxoWellcome R&D, Park Road, Ware, Herts, UK SG12 0DP*
[d] *Biomet Division, GlaxoWellcome Inc, 5 Moore Drive, Research Triangle Park, NC 27709, USA*

The applicability of novel NMR flow probe technology has been tested by the measurement of 300 MHz ^{1}H NMR spectra of a series of rat urine samples. Compared with conventional automatic operation, the method resulted in a significantly increased rate of sample throughput, required minimal spectrometer optimisation before each measurement and avoided the need for expensive and fragile NMR sample tubes. The NMR approach has been coupled with computer methods for spectral data reduction and classification using, in this case, principal components analysis. The flow probe NMR approach offers distinct advantages in situations where large numbers of samples require NMR analysis in a short period of time. These could include routine samples from high throughput chemical synthesis, biofluid samples for drug toxicity monitoring as shown here, samples for clinical diagnosis or real-time analysis in chemical production facilities.

Recently, there have been fundamental changes in the basic strategies and approaches used by the pharmaceutical industry in drug discovery. Sequential chemical synthesis is giving way to array and combinatorial methods which result in much greater numbers of samples for molecular structure and purity analysis. In addition, the increase in the numbers of drug candidate compounds presented for biological testing has also resulted in the need for the development of high throughput screens of potential toxicity. NMR spectroscopy of biofluids has been shown to provide important biochemical information relating to drug toxicity[1] but, in order to apply NMR technology to high throughput screening, further advances in the automation of high-resolution NMR spectroscopy are necessary. When coupled with the high costs of skilled personnel, the need for full time operation of high capital cost equipment and the necessity of greater experimental reproducibility, changes in operating practice are required. This is despite the fact that many NMR spectrometers are now equipped with automatic sample changing robots and associated software which facilitates sample changing, spectral parameter- and field-homogeneity optimisation, and collection and processing of data.

Conventional high-resolution NMR spectroscopy relies on the use of precision glass NMR tubes which are both delicate and expensive. Sample preparation obviously involves filling the tubes and this together with the time taken to exchange them using a conventional robotic sample tube changer limits increases in speed and efficiency for high sample throughput. An alternative approach is to use a flow probe in which direct transfer of a sample is possible from a reservoir into the NMR detector cell itself. A closely related technology using flow probes is directly coupled HPLC–NMR spectroscopy in which the output from a chromatographic column is fed to a flow probe.[2,3] In this investigation we report the use of a novel flow injection NMR detection system linked to an automatic sample-handling device in which samples are pipetted from a 96-well plate. This type of technology has only been reported recently in the scientific literature as an abstract of a meeting presentation.[4] We have used this system to measure ^{1}H NMR spectra of rat urine, from animals dosed with the model hepatotoxic drug thioacetamide, in order to monitor the altered biochemical profile of the urine as a consequence of the toxic insult. This alteration to the biochemical profile has been followed both by visual examination of the NMR spectra and by the use of principal components (PC) analysis of NMR spectral descriptors. Both approaches demonstrate changes in the levels of a number of endogenous metabolites which can be related to the toxic insult.[5,6]

Experimental

Urine samples were taken from a larger study of liver toxicity which will be reported elsewhere. For the current investigation, urine samples were taken from male Wistar rats that had been dosed orally with thioacetamide at 200 mg kg^{-1} body weight in sterile water. In all, 27 urine samples were collected, comprising 9 samples taken from control rats dosed only with sterile water and samples from two rats dosed orally with thioacetamide. Time points for urine collection were predose, 0–7 h, 7–24 h, 24–31 h, 48–55 h, 72–79 h, 96–103 h, 120–127 h and 144–151 h after dosing. The urine samples were frozen immediately after collection and thawed prior to analysis. A 950 µl aliquot of each sample was placed in a separate well of a 96-well plate and 50 µl of D_2O was added to each sample to provide a field-frequency lock. The plate was then covered with a thin sheet of paraffin wax film.

NMR spectra were measured in the stop-flow mode using a Bruker (Rheinstetten, Germany) DPX-300 instrument operating at 300.13 MHz for ^{1}H observation using a 5 mm single cell ^{1}H/^{13}C inverse detection flow probe with an active volume of 250 µl. Sample transfer from the 96-well plate to the NMR flow probe used a Gilson (Middleton, WI, USA) XL233 automatic sample handling system interfaced to the NMR data system for control and timing. For each sample, 480 µl of urine was

† Present address: Stine-Haskell Research Center, Dupont-Merck, Elkton Road, Newark, DE 19714, USA.

pipetted from the well into the transfer line and separated from subsequent samples by 500 μl of a wash solution of water, each solution being separated by an air bubble. The transfer time from the sample well to the NMR probe was approximately 45 s. A diagram showing the experimental arrangement is shown in Fig. 1.

Spectra were acquired using the NOESYPRESAT pulse sequence (Bruker) to suppress the large NMR peak from the solvent water. For each sample, 64 transients were collected into 32768 time domain data points with a spectral width of 3140.7 Hz, an acquisition time of 5.22 s and a total recycle time of 6.22 s. The FIDs were multiplied by a line-broadening function of 1.0 Hz to improve the signal-to-noise ratio and, after zero filling by an equal number of data points, were Fourier transformed, phased and baseline corrected. Chemical shifts were referenced to the methyl resonance of creatinine at δ3.05. The magnetic field was optimised (shimmed) for the first sample only and then not adjusted further during the data collection on the subsequent 26 samples. No significant loss of resolution was observed.

Each spectrum (δ10.0–δ0.24) was also segmented into 256 equal chemical shift regions using AMIX software (version 2.1.3, Bruker) and the total integrated intensity in each region was determined to provide a series of descriptors of the spectra normalised to the total integral of each spectrum to remove concentration effects. These data were autoscaled to give a mean of zero and a variance of ±1 for each descriptor and subjected to PC analysis using the software package PIROUETTE (version 2.03, Infometrix Inc, Seattle, USA) running on an IBM-compatible personal computer.

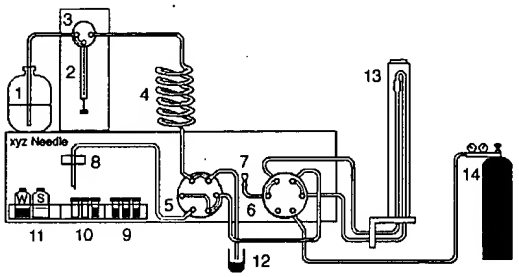


Fig. 1 Schematic representation of the automatic injector and flow probe system. 1, transport liquid reservoir; 2, sample dilutor syringe; 3, three-way valve for dilution liquid; 4, sample loop; 5, six-way valve for sample loading; 6, six-way valve for sample injection to probe; 7, injection port; 8, needle; 9, rack for sample vials or 96 well plate; 10, rack or 96 well plate for recovered samples; 11, washing fluids and waste reservoir; 12, external waste reservoir; 13, NMR flow probe; and 14, inert gas cylinder for drying.

Results

The experimental arrangement was tested using both control rat urine samples and those from animals that had received toxic doses of thioacetamide.[7] A typical 300 MHz ^{1}H NMR spectrum of urine from a control rat is shown in Fig. 2(A). Many of the endogenous species in urine have been assigned previously and these are marked on the figure.[1] The lack of cross-contamination between samples had been tested previously and it was determined that with the wash procedure given above no NMR peaks could be detected from the previous sample at the signal-to-noise ratio obtained after 64 transients (M. Spraul and M.Hofmann, unpublished results).

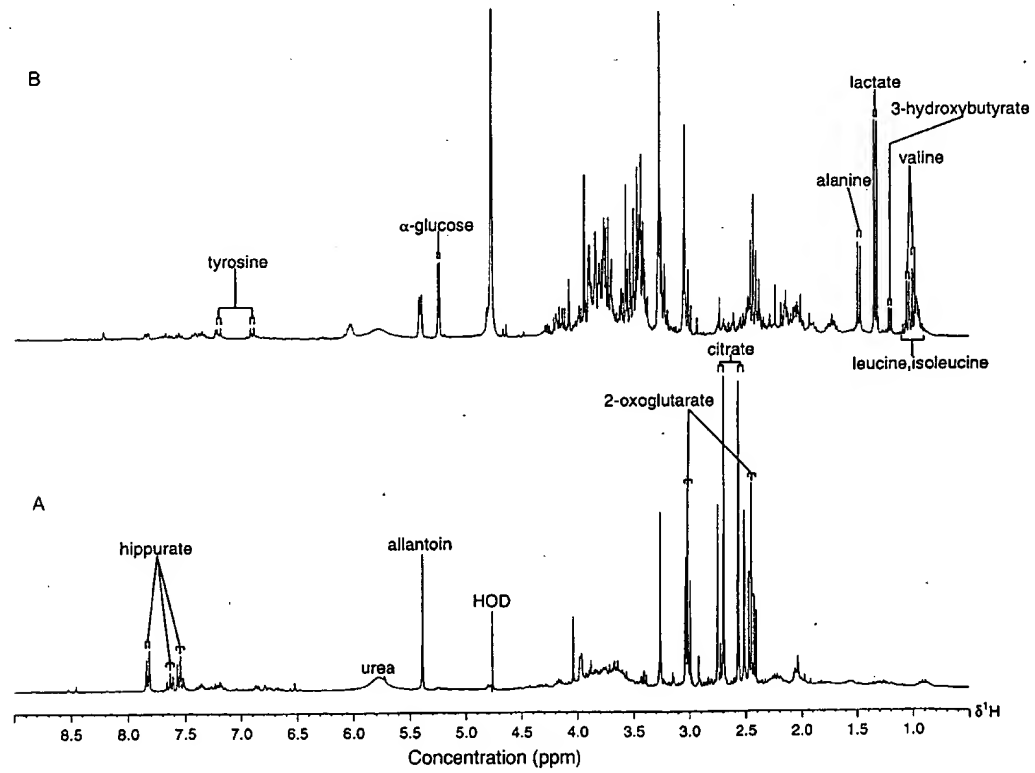


Fig. 2 ^{1}H NMR spectra of, A, control rat urine and B, rat urine 31–55 h after oral dosing with thioacetamide (200 mg kg^{-1}). Assignments are as marked.

A typical spectrum obtained from a rat urine for the period 31–55 h after administration of thioacetamide at a dose of 200 mg kg^{-1} is shown in Fig. 2(B). This time period was chosen to ensure that any metabolites of thioacetamide had already been eliminated. Despite the lack of field homogeneity adjustment, the spectral resolution remained suitable for analysis. A number of major changes can be observed including the loss of 2-oxoglutarate, succinate, citrate, and increases in lactate, alanine, α-glucose, β-glucose, 3-hydroxybutyrate, isoleucine, leucine, valine, and tyrosine.

The spectra measured in this study were segmented to produce 256 descriptors of the spectral intensity and these were used as input to PC analysis. The 0–7 h and 7–24 h time-point samples from the thioacetamide-treated animals were excluded since these were where the metabolites of thioacetamide were excreted and ^{1}H NMR resonances from these metabolites would have affected the analysis. The first two PCs accounted for 61% of the data variance and a plot of the PC scores where each point represents a urine sample is shown in Fig. 3. The PC plot indicated a distinct biochemical trajectory for the toxicology of thioacetamide through time. This was observed in the plot for urines obtained after thioacetamide dosing by a decrease in value of PC1 as the time after dosing increased. From this plot there was a return to normal, control, region of the PC plot after 55 h post-dose by a recovery trajectory similar to that observed

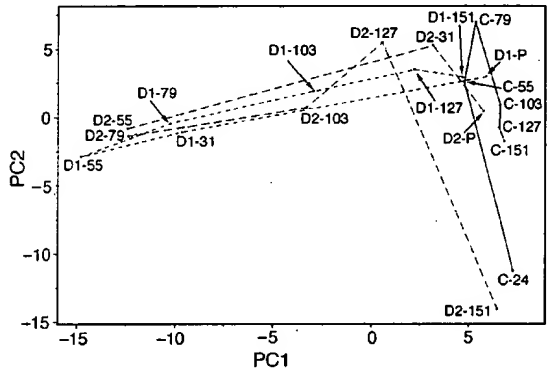


Fig. 3 Plot of the first two principal components (PC1 *versus* PC2) using descriptors taken from the NMR spectra of urine from control and dosed animals. Each point represents a separate urine sample. C, control samples and D1, D2 urine samples after dosing with thioacetamide, the second number represents the time in hours after dosing.

for the toxicity trajectory. This indication of recovery was also seen in the NMR spectra with the NMR spectral profile of the urine returning to normal by the final time point of the study. These findings are consistent with the known effects of thioacetamide which causes both centrilobular necrosis of the liver and damage to the S3 region of the kidney.[7]

Each urine sample required only approximately 3 min for NMR data collection and this resulted in more than a factor of two increase in sample throughput as compared with a conventional autosampler using NMR glass tubes. This arose mainly because of the lack of need to optimise field homogeneity for each sample. Further substantial increases in sample throughput will be possible through the use of increased sample volumes or decreased NMR data acquisition times, leading to an estimated total requirement of 1 min for each ^{1}H NMR analysis. Two-dimensional ^{1}H–^{1}H correlation NMR spectra using magnetic field gradients for coherence selection then become possible using only approximately 5 min data acquisition time, leading to further possibilities of high throughput NMR/pattern recognition classification studies based on such spectra.

In summary therefore, the generation of chemical structural or biochemical information from ^{1}H NMR spectroscopy of samples using high throughput flow-probe technology promises to provide new and valuable tools for rapid chemical analysis and, when coupled to pattern recognition classification methods, it will be of importance for biological screening of candidate drug compounds.

References

1 Nicholson, J. K., and Wilson, I. D., *Prog. Nucl. Magn. Reson. Spectrosc.*, 1989, **21**, 449.
2 Lindon, J. C., Nicholson, J. K., and Wilson, I. D., *Adv. Chromatogr. (N.Y.)*, 1995, **36**, 315.
3 Lindon, J. C., Nicholson, J. K., and Wilson, I. D., *Prog. Nucl. Magn. Reson. Spectrosc.*, 1996, **29**, 1.
4 Keifer, P. A., *Abs. Pap. ACS*, 1997, **213**, 277.
5 Gartland, K. P. R., Beddell, C. R., Lindon, J. C., and Nicholson, J. K., *Mol. Pharmacol.*, 1991, **39**, 629.
6 Holmes, E., Bonner, F. W., Sweatman, B. C., Lindon, J. C., Beddell, C. R., Rahr, E., and Nicholson, J. K., *Mol. Pharmacol.*, 1992, **42**, 922.
7 Waterfield, C. J., Turton, J. A., Scales, M. D. C., and Timbrell, J. A., *Arch. Toxicol.*, 1993, **67**, 244.

Paper 7/05551J
Received July 31, 1997
Accepted September 29, 1997

Mellerson, Kendra

From: Gakh, Yelena
Sent: Tuesday, August 05, 2003 2:33 PM
To: STIC-EIC1700
Subject: 09890973

Dear Kendra:

please order one more list:

13. TITLE: "Plant histochemistry by correlation peak imaging"
AUTHOR(S): *Metzler, A.; Izquierdo, M.; Ziegler, A.; Koeckenberger, W.; Komor, E.; von Kienlin, M.; Haase, A.; Decorps, M.*
CORPORATE SOURCE: Physikalisches Inst. V, Univ. Wuerzburg, Wuerzburg, 97074, Germany
SOURCE: **Proceedings of the National Academy of Sciences of the United States of America (1995), 92 (25), 11912-15**

Thank you,

Yelena

Yelena G. Gakh, Ph.D.

Patent Examiner
USPTO, cp3/7B-08
(703)306-5906

Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 11912–11915, December 1995
Biophysics

Plant histochemistry by correlation peak imaging

(plant physiology/nuclear magnetic resonance/*Ricinus communis* seedlings)

A. METZLER*, M. IZQUIERDO†, A. ZIEGLER†, W. KÖCKENBERGER‡, E. KOMOR‡, M. VON KIENLIN*§, A. HAASE*, AND M. DÉCORPS†

*Physikalisches Institut V, Universität Würzburg, Am Hubland, 97074 Würzburg, Germany; †Institut National de la Santé et de la Recherche Médicale, Unité 438, Université J. Fourier, Centre Hospitalier Universitaire, BP 217, 38043 Grenoble, France; and ‡Botanisches Institut (Pflanzenphysiologie), Universität Bayreuth, Universitätsstrasse 30, 95440 Bayreuth, Germany

Communicated by Richard R. Ernst, Eidgenossische Technische Hochschule Zentrum, Zurich, Switzerland, August 29, 1995

ABSTRACT **Using a new NMR correlation-peak imaging technique, we were able to investigate noninvasively the spatial distribution of carbohydrates and amino acids in the hypocotyl of castor bean seedlings. In addition to the expected high sucrose concentration in the phloem area of the vascular bundles, we could also observe high levels of sucrose in the cortex parenchyma, but low levels in the pith parenchyma. In contrast, the glucose concentration was found to be lower in the cortex parenchyma than in the pith parenchyma. Glutamine and/or glutamate was detected in the cortex parenchyma and in the vascular bundles. Lysine and arginine were mainly visible in the vascular bundles, whereas valine was observed in the cortex parenchyma, but not in the vascular bundles. Although the physiological significance of these metabolite distribution patterns is not known, they demonstrate the potential of spectroscopic NMR imaging to study noninvasively the physiology and spatial metabolic heterogeneity of living plants.**

In the tissue of plant organs, enzymatic reactions and metabolic pathways are compartmentalized. A striking example is C4-photosynthesis, where the different reaction steps are spatially separated between mesophyll and bundle sheath cells (1). However, the knowledge about the distribution and the concentration of metabolites in plants is still very limited, mainly because of the lack of appropriate experimental techniques. Only a few methods are available to study the localization of metabolites in plant materials. Enzyme localization is accessible by methods of molecular biology—for example, by cDNA *in situ* hybridization and immunohistochemistry, by tissue print (2), or by measuring the activity of β-glucuronidase (3). Extraction of cell sap by microcapillaries is possible only from relatively large cells located close to the surface of the plants (4). Fixation procedures in microautoradiography (5) and electron-dispersive energy-loss spectroscopy (6) of water-soluble compounds or elements might disturb the spatial distribution of metabolites. All of these methods have in common an invasive or even destructive way of measuring the spatial distribution of the constituents of the tissue.

Nuclear magnetic resonance (NMR) measurements are noninvasive by nature. NMR imaging, which is based on the NMR signals from the hydrogen in water molecules, has had an enormous impact on medical diagnostics by visualizing human anatomy in great detail. NMR spectroscopy can provide information on the different chemical constituents in a sample by detecting slight shifts of their resonance frequencies and is widely used in analytical chemistry. The combination of NMR imaging and spectroscopy resulted in a technique known as "chemical-shift imaging (CSI)" (7, 8). CSI enables the spatial distribution of specific chemical compounds within a heterogeneous sample to be measured. Initial applications of CSI to plants have already provided some insight into the spatial distribution of metabolites (9, 10). Being inherently noninvasive, these NMR measurements fully preserve the integrity of the plant. They affect neither its physiology nor the concentrations of the metabolites *in situ*. Therefore, NMR imaging and spectroscopy applied to study plant materials may yield valuable information that cannot be obtained by using any conventional, destructive method.

In data acquired by normal CSI with one spectral dimension, it is sometimes impossible to differentiate between components with overlapping resonance lines. This problem arises particularly in ^{1}H-NMR spectroscopy with its inherently limited spectral dispersion. By using two-dimensional (2-D) correlation NMR spectroscopy (11), the spectral resolution and consequently the information content of the spectra can be improved considerably. Correlation spectroscopy and other multidimensional spectroscopic techniques are already a standard tool in analytical chemistry and in the study of protein structure. First *in vivo* applications of correlation spectroscopy in animals were reported recently (12, 13). In these experiments, specific molecules are identified by their characteristic correlation peaks (i.e., their off-diagonal resonances in a 2-D frequency map, indicating scalar coupled spins within the molecule). Fig. 1 shows a two-dimensional correlation map obtained *in situ* in a plant seedling and demonstrates the wealth of information available with this technique. A large number of chemical constituents including sugars and amino acids and even various anomers can be observed, representing the average concentration of these substances in the examined cross-section of the stem.

For further localization within the plant, we have added phase-encoding gradients to correlation spectroscopy (14) to obtain a correlation-peak imaging (CPI) experiment with two spatial and two spectral dimensions.¶ From the acquired data, a complete 2-D correlation map can be reconstructed for each volume element, showing the metabolite pattern at that location. Furthermore, the spatial distribution of specific metabolites can be visualized by displaying the spatially varying intensity of the corresponding correlation peaks. These "metabolic images" represent the distribution of the metabolites, with the assumption of uniform metabolite relaxation times in all plant tissues. Conventional ^{1}H-NMR images of the plant with high spatial resolution can be acquired in the same experimental setup and allow the

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: CPI, correlation-peak imaging; 2-D, two dimensional; CSI, chemical-shift imaging.

§To whom reprint requests should be addressed.

¶Metzler, A., Izquierdo, M., Ziegler, A., Lefur, Y., Köckenberger, W., Komor, E., von Kienlin, M., Haase, A. & Décorps, M., Proceedings of the 11th Annual Meeting of the European Society for Magnetic Resonance in Medicine and Biology, April 20–24, 1994, Vienna, Austria, p. 491 (abstr.).

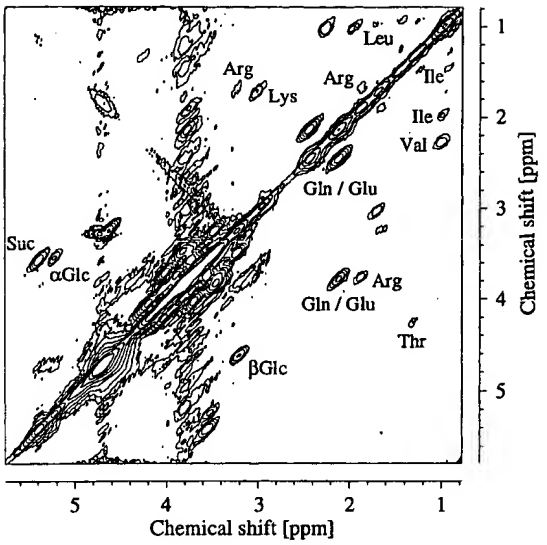


FIG. 1. NMR 2-D correlation spectrum obtained from a 4-mm slice selected *in situ* in the hypocotyl of a 6-day-old castor bean seedling. Most interesting are the spots appearing off the diagonal: the positions of these correlation peaks are characteristic of specific molecules and originate from spins presenting scalar couplings to neighboring spins in the same molecule. From their position, the correlation peaks can be assigned to specific substances (Suc, sucrose; Glc, glucose; and amino acids indicated by their standard three-letter code); even two anomers of glucose can be distinguished. This is the global spectrum of the slice selected in the hypocotyl without further localization, representing the average amount of the detected metabolites in this volume. The goal of the CPI experiment is to measure the spatial distribution of these substances within the slice.

correlation of the measured metabolic distributions with the anatomy of the plant.

METHODS

One of our first attempts to demonstrate the potential of the CPI technique was to measure the spatial distribution of the most abundant carbohydrates and amino acids (sucrose, α- and β-glucose, glutamine/glutamate, arginine, lysine, and valine) in the hypocotyl of a 6-day-old castor bean seedling (*Ricinus Communis L.*) (16). Seedlings were grown in darkness on top of glass tubes fitted into a standard microimaging NMR probe. Thus, the plant could be placed into the spectrometer without disturbing its physiological environment. All experiments were performed on a Bruker (Karlsruhe, Germany) model AMX500 NMR spectrometer, equipped with an 89-mm bore, 11.75-T superconducting magnet, and a shielded imaging gradient system. Both ^{1}H-^{1}H-CPI experiments and conventional NMR imaging experiments with high spatial resolution were conducted for every plant.

RESULTS

The results for one plant are shown in Figs. 2 and 3. In the high-resolution proton image of the hypocotyl anatomy (Fig. 2), eight vascular bundles, the pith parenchyma, and the cortex parenchyma can be seen. The phloem and the xylem, which are important for sucrose and water transport, respectively, can be clearly distinguished within the vascular bundles. Experimental parameters for this microscopic NMR image with a nominal spatial resolution of 24 μm are given in the figure caption.

The metabolite images obtained with the CPI experiment are presented in Fig. 3. They show the distribution of sucrose,

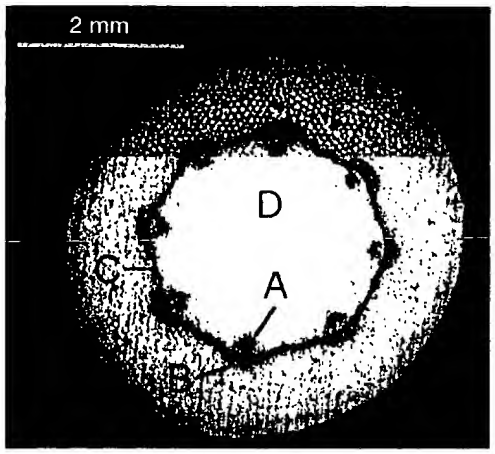


FIG. 2. High-resolution proton NMR image of a cross section of the hypocotyl, which allows the anatomy of the plant to be identified in great detail. Each of the eight vascular bundles consists of the xylem region (A) at the inner side and the phloem region (B) at the outer side of the meristem ring (C). The cellular structure of pith (D) and cortex parenchyma (E) is clearly visible. Cell wall material appears dark. This inversion recovery spin echo image was acquired in 49 min with an inversion delay of 750 msec, an echo time of 8 msec, and a repetition time of 5.75 sec. The 256 × 256 image matrix with a field of view of 6 mm × 6 mm and a slice thickness of 1 mm resulted in a nominal in-plane resolution of 24 μm × 24 μm.

of glucose, and of some amino acids, which can be correlated to the anatomy of the plant by superimposing the metabolite images and the high resolution image in Fig. 2. Since sucrose is the dominant carbohydrate in the phloem, we expected and found high sucrose concentrations in the vascular bundles (Fig. 3*A*). However, the two stereoisomers of glucose were mainly found in the pith parenchyma (α- and β-glucose; Fig. 3 *B* and *C*). The observation that the cortex parenchyma is rich in sucrose, whereas the pith parenchyma is rich in glucose, was unexpected. The biological significance of this complementary spatial distribution must be speculative at this early stage: the prevalence of hexoses in the pith parenchyma might contribute to a sufficiently high osmotic potential serving to maintain the turgor of the hypocotyl. The different locations of sucrose and glucose may have important consequences for the conflicting models of extension growth of shoots.

Glutamine is the major amino acid in the phloem sap (17) and is considered to be the main nitrogen carrier in castor bean seedlings. In the metabolite images, glutamine/glutamate occurs mostly in the cortex and the vascular bundles (Fig. 3*E*), whereas lysine (Fig. 3*F*) and arginine (Fig. 3*G*) are prevalent in the vascular bundles only. In earlier studies analyzing phloem exudate, it was found that the arginine concentration in the sieve tubes is not higher than in extracts of hypocotyl tissue (17). However, our CPI results reveal a prevalence of arginine in the vascular bundles. This may indicate an enrichment of arginine in the bundle parenchyma cells. The metabolic image of valine (Fig. 3*H*) shows a distribution that is restricted to the cortex parenchyma outside the vascular bundles. Within our limit of sensitivity, we could not observe any valine cross-peak in the vascular bundles.

DISCUSSION

The CPI experiment enables the observation of molecules that are typically accessible by NMR spectroscopy in the liquid state. These molecules must be relatively small and mobile, because any motional restriction of the spins results in a

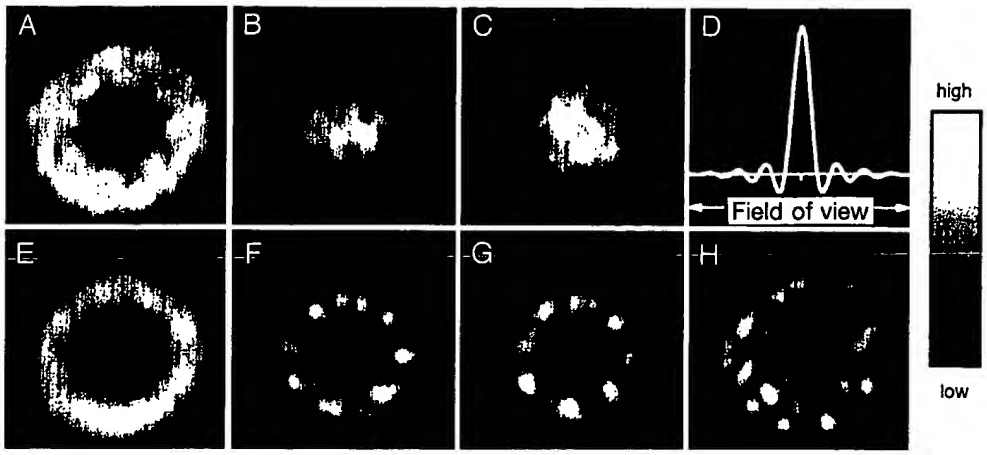


FIG. 3. Metabolic images obtained in a CPI experiment, in the same cross section of the hypocotyl as shown in the NMR image in Fig. 2. These images were obtained by selecting individual cross-peaks in the correlation spectra and by displaying their spatial distribution. The gray scale was adjusted individually for each image to obtain maximal contrast. (*A*) The distribution of the sucrose cross-peak shows high intensity in the vascular bundles, while the signal is lower in the cortex parenchyma. In the pith parenchyma, the sucrose intensity was low and decreasing towards the center, confirming the results of earlier CSI experiments (10). Because of the gray scaling, this gradient cannot be seen in *A*, but it clearly appears when plotting intensity profiles. The distribution of the signals corresponding to α- and β-glucose (*B* and *C*, respectively) reveals high intensity in the pith parenchyma. The signal of glutamine/glutamate (*E*) appears as a bright ring covering the cortex parenchyma and the vascular bundles. Despite their low concentration, lysine (*F*) and arginine (*G*) can be observed in the vascular bundles. Valine (*H*) was only found in the cortex parenchyma but not (within the limits of sensitivity) in the vascular bundles. Experimental details: in the four-dimensional CPI experiment, 16 × 16 localized correlation spectra were acquired in a field of view of 6 mm × 6 mm. The selection of a 4-mm slice along the hypocotyl resulted in a nominal volume of 560 nl for each image voxel. After Fourier transformation, the integrated intensity of individual cross-peaks was extracted, Fourier-interpolated to obtain a 256 × 256 image matrix, and scaled individually. The quality of the spatial localization can be assessed by the point spread function shown in *D*.

broadening of the resonance frequencies and a shortening of the transverse relaxation time, thereby impeding detection in the CPI experiment. This, in turn, will hinder the detection of molecules that are bound to membranes and may also reduce the "NMR visibility" of compounds fixed in larger storage molecules. Another limitation of NMR spectroscopy is its inherently low sensitivity. The lower limit whereby concentration can still be detected is determined—among other parameters—by the strength of the main magnetic field of the spectrometer, by the duration of the experiment, and by the size of the voxels in the metabolic image—i.e., the spatial resolution. The higher the magnetic field or the longer the experiment or the larger the voxels, the lower is the detectable concentration limit. In our experiments at 11.7 Tesla, with an experimental duration of 4 h and 33 min and a voxel size of 560 nl, we were able to observe metabolite concentrations to the order of 10 mM. Increasing the experimental duration may be used to improve spatial resolution or to increase the sensitivity of the CPI experiment to detect lower concentrated metabolites. Finally, the size of the experimental arrangement including the regulation of environmental parameters such as temperature, humidity, and light is limited by the space available in the NMR instrument; the bore size of our instrument was 89 mm.

The CPI experiment makes possible the identification of a large variety of chemical compounds and the measurement of their spatial distribution within the plant. The CPI experiment even enables one to distinguish between various stereoisomers. It has been shown that transport and metabolic reactions can depend strongly on the stereospecificity (15, 18), but until now the stereo configuration of basic sugars in the plant cells has not been known. In the same experimental setup, NMR images with high spatial resolution can be obtained, which allow one to correlate the measured distribution of metabolites with the anatomy of the plant. The main advantage of NMR methodology lies in its noninvasiveness. The experiments may be conducted repeatedly on the same plant and can monitor dynamic changes of the metabolites, typically in response to changing experimental parameters. For instance, the osmoregulation of cell turgor by changing hexose concentrations could be continuously monitored on an individual plant. CPI could thus become a versatile tool in studies of the reaction of plants to environmental stress. The combination of the molecular information accessible by multidimensional spectroscopic techniques and of the spatial information obtained by CSI may open a wide field of research in plant physiology.

We particularly thank Dr. R. Bligny, Dr. R. Massarelli, Dr. G. Orlich, Prof. J. N. Kanfer, and Prof. U. Heber for their encouragement and stimulating discussions. This study was financially supported by grants received from the Deutsche Forschungsgemeinschaft (A.M., Ha 1232/11-1; W.K., SFB 137; M.v.K., Ki 433/2-1; A.H., Ha 1232/8-1) and—within the PROCOPE framework—from the Deutscher Akademischer Austauschdienst and from the French Ministère des Affaires Etrangères.

1. Hatch, M. D. & Osmond, B. (1976) in *Encyclopedia of Plant Physiology: New Series*, eds. Stocking, C. R. & Heber, U. (Springer, Heidelberg), Vol. 3, pp. 144–177.
2. Zheng-Hua, Y. & Varner, J. E. (1991) *Plant Cell* **3**, 23–27.
3. Yang, N. S. & Russel, D. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 4144–4148.
4. Fricke, W., Leigh, R. A. & Tomos, A. D. (1994) *Planta* **192**, 310–316.
5. Fritz, E. (1980) *Ber. Dtsch. Bot. Ges.* **93**, 109–121.
6. Probst, W. & Bauer, R. (1987) *Verh. Dtsch. Zool. Ges.* **80**, 119–123.
7. Kumar, A., Welti, D. & Ernst, R. R. (1975) *J. Magn. Reson.* **18**, 69–83.
8. Brown, T. R., Kincaid, B. M. & Ugurbil, K. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 3523–3526.
9. Rumpel, J. & Pope, J. M. (1992) *Magn. Reson. Imaging* **10**, 187–194.
10. Metzler, A., Köckenberger, W., von Kienlin, M., Komor, E. & Haase, A. (1994) *J. Magn. Reson. Ser. B* **105**, 249–252.
11. Ernst, R. R., Bodenhausen, G. & Wokaun, A. (1987) *Principles of Nuclear Magnetic Resonance in One and Two Dimensions* (Oxford Univ. Press, Oxford).
12. Berkowitz, B. A. (1992) in *NMR 27 In-Vivo Magnetic Resonance Spectroscopy II: Localization and Spectral Editing*, eds. Diehl, P.,

Fluck, E., Günther, H., Kosfeld, R. & Seelig, J. (Springer, Heidelberg), pp. 223–236.
13. Ziegler, A., Izquierdo, M., Rémy, C. & Décorps, M. (1995) *J. Magn. Reson. Ser. B* **107**, 10–18.
14. Cohen, Y., Chang, L.-H., Litt, L. & James, T. L. (1989) *J. Magn. Reson.* **85**, 203–208.
15. Ehwald, R., Sammler, P. & Göring, H. (1973) *Folia Microbiol. (Prague)* **18**, 102–117.
16. Komor, E., Orlich, G., Köhler, J., Hall, J. L. & Williams, L. E. (1991) in *Recent Advances in Phloem Transport and Assimilate Compartmentation*, eds. Bonnemain, J. L., Delrot, S., Lucas, W. J. & Dainty, J. (OUEST Editions, Nantes, France), pp. 301–308.
17. Schobert, C. & Komor, E. (1989) *Planta* **177**, 342–349.
18. Komor, E., Schobert, C. & Cho, B.-H. (1985) *Eur. J. Biochem.* **146**, 649–656.

Mellerson, Kendra

From: Gakh, Yelena
Sent: Tuesday, August 05, 2003 2:33 PM
To: STIC-EIC1700
Subject: 09890973

Dear Kendra:

please order one more list:

14. TITLE: "Using multivariate methods on solid-state 13C NMR data of complex materials"
AUTHOR(S): *Karlstroem, Hans; Nilsson, Mats; Norden, Bo*
CORPORATE SOURCE: Kimit AB, Kiruna, S-981 86, Swed.
SOURCE: **Analytica Chimica Acta (1995), 315(1-2), 1-14**

Biotech QD71. A47
or Adams

Thank you,

Yelena

Yelena G. Gakh, Ph.D.

Patent Examiner
USPTO, cp3/7B-08
(703)306-5906

Article title	Using multivariate methods on solid-state 13C NMR data of complex materials
Article identifier	0003267095107018
Authors	Karlstrom_H Nilsson_M Norden_B
Journal title	Analytica Chimica Acta
ISSN	0003-2670
Publisher	Elsevier Netherlands
Year of publication	1995
Volume	315
Issue	1-2
Supplement	0
Page range	1-14
Number of pages	14
User name	Adonis
Cost centre	Development
PCC	$20.00
Date and time	Tuesday, August 05, 2003 4:31:35 PM

ELSEVIER

Analytica Chimica Acta 315 (1995) 1–14

ANALYTICA CHIMICA ACTA

Using multivariate methods on solid-state ^{13}C NMR data of complex materials

Hans Karlström [a,*], Mats Nilsson [b], Bo Nordén [c]

[a] *Kimit AB, S-981 86 Kiruna, Sweden*
[b] *Department of Forest Site Research, Swedish University of Agricultural Sciences, S-901 83 Umeå, Sweden*
[c] *Medicinal Chemistry II, Computational Chemistry Group, Astra Hässle AB, S-431 83 Mölndal, Sweden*

Received 14 November 1994; accepted 27 April 1995

Abstract

Multivariate data analysis (MVA) has been used as an aid in the analysis and interpretation of ^{13}C NMR spectra in the solid state. The goal of this study was to investigate the effect of some important instrumental parameters and calculation strategies on the outcome of the multivariate data analysis. The samples used were two peat forming plants, *Sphagnum fuscum* and *Carex rostrata*, incubated in four different redox environments. It was found that normalising each NMR spectrum to a constant area should be avoided. Using non-normalised data we get a slightly better class separation and the peaks in the 'subspectra' are sharpened. Depending on the relative size of interesting variation one should be careful when choosing the number of variables, i.e. number of data points characterising each spectrum. The line broadening technique should be used with great care in order not to obscure the information. We also suggest the use of the free induction decay (FID)/MVA directly for classification purposes. This is a new approach to analyse the output data from NMR measurements.

Keywords: Nuclear magnetic resonance spectrometry; Peat forming plants; Multivariate data analysis; Principal component analysis; Free induction decay

1. Introduction

The use of ^{13}C CP/MAS NMR on complex and heterogeneous material is widely used in many research areas, e.g. wood and pulp chemistry, soil and humus research, peat science and decomposition studies [1–8]. The advantages of using solid-state ^{13}C NMR are several, viz. it is a non-destructive method, it gives a good picture of the distribution of carbon atoms in different chemical compounds and the method does not require any extraction procedures or other pretreatments, except maybe drying and milling. Since no chemical pretreatment is needed one expects that the chemical compounds in the sample are not changed in any way and that the result from the analysis should mirror as close as possible the native sample. The disadvantage is the low sensitivity of the method, which comes from the

* Corresponding author.

SSDI 0003-2670(95)00235-9

low natural abundance of carbon-13 (1.108%), the gyromagnetic constant of the carbon nuclei (four times lower than of the proton) and the slow carbon relaxation rates. In order to get a reasonable S/N ratio in ^{13}C solid-state NMR analysis, the number of scans has to be increased. This means that the total analysis time will be rather long (2–5 hours) depending on the type and complexity of the sample.

Solid-state carbon-13 NMR spectra of complex materials often consist of broad overlapping peaks which make the interpretation somewhat complicated. However, there are several methods to solve the multicomponent spectrum, e.g. simulated/experimental spectra of smaller molecules that are supposed to create the overall spectrum. Another approach is to use multivariate techniques such as PCA (principal component analysis) and PLS (partial least squares) [9–12]. By using multivariate data analysis it is possible to extract principal components from a set of complex spectra and display 'subspectra' from the loadings, which hold the information of the variables.

Multivariate data analysis methods are presently used in many fields of chemical research [13–15]. The use of solid-state carbon-13 NMR in combination with multivariate data analysis is not very abundant in the literature [16–19] even though the combination should be very powerful in the analysis and interpretation of crowded and non-resolved spectra, such as those produced by heterogeneous materials. In solution NMR there are several examples of combining NMR and multivariate methods [20–26], where usually the chemical shift value is used as input data for multivariate data analysis. In solid-state NMR the whole spectrum can be digitised and the amplitude of the signal at certain frequencies can be used as input data. In this manner the data contain information both of the chemical shift and the relative intensities.

The aim of this paper is to investigate some important parametric aspects of using solid-state carbon-13 NMR data in combination with multivariate data analysis.

Normalising data is a common procedure, but is it necessary and how is the multivariate result effected by this pretreatment? The number of variables (descriptors) for each sample is very big when spectroscopic methods are used. Often the number of variables is reduced in order to speed up the calculation, this is justified by the fact that neighbouring data points are very covariant if the spectrum is smooth. Depending on the relative differences within the data one can reduce the number of variables. In NMR spectroscopy one can increase the S/N ratio by applying an exponential decaying function to the FID (free induction decay) before Fourier transformation, but how will this affect the result of the multivariate data analysis? The last question treated in this paper deals with the use of FID data instead of spectrum data.

2. Materials and methods

2.1. Decomposition experiment

Two of the most common peat forming plants in the northern part of the northern hemisphere (the moss *Sphagnum fuscum* and the sedge *Carex rostrata*) were chosen as substrates for the decomposition experiments. An additional substrate has been used, viz. a 1:1 mixture of *Sphagnum fuscum* and *Carex rostrata*. The plant material was collected in September. For *Carex rostrata* only the vegetative part above ground biomass was used. For *Sphagnum fuscum* one-year-old parts were used, collected approximately 3–6 cm below the capitula. Each of the substrate types was incubated in four different redox conditions: A, air; B, nitrogen; C, non-flushing nitrogen; and D, alternating A and B every two weeks. Experiments A, B and D were automatically flushed (100 ml/min) and shaken during 15 minutes every 12 hours, to ensure oxygenation and/or removal of produced volatile compounds. Experiment C was incubated without any flushing. All the incubation experiments were performed in duplicate at 16°C in darkness. The plant material for experiments A, B and D (15 g, dry weight) were placed in 1000-ml glass bottles together with water giving a final volume of 800 ml and a head space of 200 ml. Experiment C was performed in 60-ml serum jars, which were evacuated and refilled with pure nitrogen and sealed with butyl rubber stoppers. The water used for the incubation was collected from the same part of the mire as the plant material and filtered (Munktell's cellulose filter No. 5, Grycksbo, Sweden). Small

Table 1
Overview of the substrates and the redox environments. Three different substrates have been incubated under four different redox conditions

Substrate	Redox condition [a]			
	Air	Nitrogen	Non-flushing nitrogen	Alternating
Sphagnum fuscum	SA	SB	SC	SD
Carex rostrata	CA	CB	CC	CD
S and C (1:1)	XA	XB	XC	XD

[a] S = *Sphagnum fuscum*, C = *Carex rostrata*, X = mixture of S and C (1:1), A = air-treated, B = nitrogen-treated, C = non-flushing nitrogen, D = two weeks with condition A and two weeks with condition B.

amounts of the plant material and the water were removed at each sampling occasion in such proportions that the ratio of water and plant material was retained throughout the incubation experiment. Approximately every second month ^{13}C NMR analysis was performed (see Tables 1 and 2). The wet sample was frozen in a freezer (−20°C) and then freeze-dried. The sample was then ground in a ball-mill in such a manner that the temperature never came above 50°C. The samples were kept in a freezer at −20°C until they were analysed with solid-state NMR.

2.2. NMR analysis

The ^{13}C CP/MAS NMR spectra were recorded with a Bruker MSL-100 at 25.178 MHz. The parameters for the ^{13}C CP/MAS experiments have been investigated earlier [7]. If not stated differently the parameters were as follows; 1 ms contact time, 2.5 s repetition delay, 5000 scans of 700 data points zero filled to 2 K, 20 Hz line broadening (LB) and 3000 Hz ± 5 Hz spinning rate using double air-bearing PSZ-rotors (NILCRA), outer diameter 7 mm, with Kel-F caps. The frequency shift scale is externally referenced to adamantane (δ-CH_2 = 38.3 ppm, 964.3 Hz, relative to tetramethylsilane).

At each sampling occasion the samples were analysed in random order, in order to avoid any introduction of systematic variation due to possible instabilities of the instrument. The FIDs were multiplicated with an exponential function (LB = 20 Hz) in order to enhance the S/N ratio and for smoothening the spectra. Instead of manually correcting the phases of the Fourier transformed FIDs all spectra were transformed to the magnitude calculated (MC) spectrum mode. This is done by adding the squares of the real and the imaginary part of the spectra and finally taking the square root of the sum.

$$MC = \left[(real)^2 + (imag)^2\right]^{1/2} \quad (1)$$

The reason for doing this is to avoid introduction of variations in the spectra due to a subjective manual phase correction. The spectral region used was ranging from 0 to 5000 Hz, corresponding to 0 to 199 ppm, covering the whole spectral area of interest. The number of data points in the original spectra (2048) was reduced to 94 or 205 equally spread over the selected area by including every 22nd and 10th data point, respectively. All data, which were digitised NMR spectra, were transferred to an IBM PC as ASCII files.

2.3. Data analysis

In order to analyse all samples in an efficient way we used multivariate data analysis. This methodology has the big advantage of being able to analyse several variables and objects simultaneously and most important, it can handle the covariance between variables. It is a fact that analyses using a spectroscopic or a chromatographic method produces a large amount of information. Firstly, there is much of information in each spectrum, which can be troublesome if the sample is complex as is the case with wood, pulp, foods, wine, beer, plants, peat or other heterogeneous mixtures producing spectra with many

Table 2
Number of the sampling occasions and the corresponding days of incubation

Sampling occasion	Days of incubation
1st	0
2nd	49
3rd	105
4th	198
5th	254
6th	345

signals that severely overlap. Secondly, the number of analysed samples introduces another problem, viz. the comparison problem. It is very difficult to do multivariate comparisons, comparing several variables and objects/samples at the same time in tables, nevertheless man has an excellent ability to analyse images, comparing classes, clusters, outliers, trends and so on.

One way to deal with these problems is to use some sort of multivariate data analysis in order to reduce the number of variables and construct pictures of the data set. This has been done in several areas with good results [14,16,20,21,27–30]. We have chosen to use the SIMCA-package [31], which includes several levels of multivariate data analysis, i.e. PCA (principal component analysis), classification and PLS (partial least squares). Regression techniques have been thoroughly described in the literature [9–12], however, a brief description will be given.

Each sample (spectrum, object) is described by a set of variables, in our case intensity of NMR signals at specific frequencies. Each new NMR analysis will generate a new object with its own set of intensities at the chosen frequencies. In the end there will be a matrix with n objects (samples) and m variables. Every object (spectrum) can be represented as a point in an m-dimensional variable space. These points can be projected to a smaller space (line, plane or hyperplane) spanned by the principal components (PC's) and describing the maximum variance within the objects. Usually the number of principal components is much lower than the original number of variables due to covariance between variables.

The procedure of principal component analysis creates two sets of vectors. The score vectors hold the information of the position of the objects in the new coordinate system called the score space. The other vectors, the loading vectors, hold the information on the relation between the new coordinate system spanned by the PC's and the coordinate system spanned by the original variables. To examine clustering, similarities and dissimilarities between the objects it is now easy to plot the calculated principal components and score vectors. Outliers are easily detected and should be examined why they are acting as outliers. By plotting the loading vectors produced by each principal component it is straightforward to see which variables are responsible for the observed behaviour in the score space. Each new principal component is perpendicular to the former and therefore independent of each other. The optimal number of principal components to be extracted is determined by cross-validation (CV) [32].

The principal component analysis, which separates the original data matrix (X) into structure (TP') and noise (E), can be expressed in mathematical terms:

$$X = 1x + TP' + E \tag{2}$$

where x is the mean vector included in the model in order to centre the objects around the mean value. The structure of the data is the product of the score matrix (T) and the loading matrix (P'). The score matrix (T) holds the information of where the objects are in the new coordinate system spanned by the principal components. The loading matrix (P') holds information of how much each variable contributes to the extracted components. E is the residual matrix left after the principal component analysis has extracted the systematic information.

A nice feature with spectroscopic data (in this case NMR data) is obtained by plotting the loading values from the loading vector of each variable against the variable number. This creates a 'subspectrum' which displays the contribution of the variables to the distribution of the objects in the score space [18]. In chemical-analytical terms it should be possible to analyse what kind of carbon signals, i.e. chemical functionalities, chemical compounds or classes of chemical compounds, which are contributing to a specific distribution in the score space. The contributions from the variables can be positively or negatively correlated to the principal component. Since the NMR spectrum is describing different kinds of carbon nucleus, the variables describe carbons in different chemical and morphological environments.

Due to scaling, centration and rotation of the data the axis of the score and loading space have no units. The score vectors are combinations of all the original variables and the loading for one variable is cos α, where α is the angle between the score vector and the variable axis. This means that the units of the scores and loadings are not easily determined based on the original units.

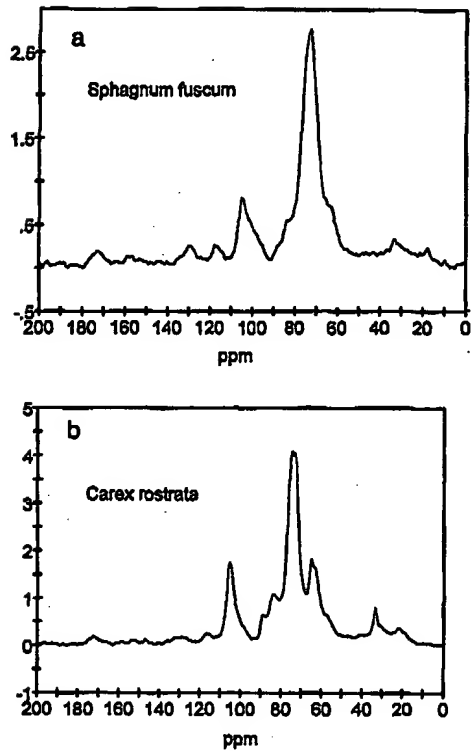


Fig. 1. (a) ^{13}C CP/MAS NMR spectrum of *Sphagnum fuscum* showing the starting material. Line broadening (LB) = 0 Hz. (b) ^{13}C CP/MAS NMR spectrum of *Carex rostrata* showing the starting material. LB = 0 Hz.

3. Results and discussion

3.1. The nature of the real data used

Two typical NMR spectra of *Sphagnum fuscum* and *Carex rostrata* are shown in Fig. 1a and 1b. The chemical shift range is from 0 ppm to 200 ppm and notice should be made of the severe signal overlaps. Assignments of the peat spectra can be found in the literature [1–3,33–37]. Usually spectra of complex materials are divided into chemical shift areas describing different functionalities of the chemical compounds. The areas are approximately as follows: 0–50 ppm, aliphatic carbons; 50–90 ppm, ring carbon of carbohydrates; 90–110 ppm; anomeric carbon of carbohydrates; 108–138 ppm olefinic and aromatic carbons; 138–160 ppm, phenolic and N-substituted aromatic carbons; 160–200 ppm carboxylic, amid and ester carbons.

In the first analysis (CALC1) 144 objects with 94 variables were used as input data in the multivariate data analysis. The objects included were *Sphagnum fuscum*, *Carex rostrata* and the mixture samples. Redox conditions (A–D) were also included. Three significant (according to cross-validation) principal components explained 90.6% of the total variance within data. Table 3 lists the results of the calculations.

A plot of the second versus the first principal component is shown in Fig. 2a. The first principal component (PC1) shows no systematic variation depending either on the botanical origin or on the degree of decomposition. The second principal component (PC2) is separating the peat classes; *Sphagnum fuscum*, *Carex rostrata* and the 1:1 mixture. The separation between *Sphagnum fuscum* and *Carex rostrata* is quite satisfying and the manually blended mixture of *Sphagnum fuscum* and *Carex rostrata* is situated in between the two pure peat/plant classes. The third principal component versus the second is shown in Fig. 2b. The third principal component (PC3) is describing the time of incubation. In the lower part of the Fig. the starting material is clustered and generally the incubation time increases when going upwards. There is a slight deviation from the trend and this is probably due to a non-linear behaviour of the decaying process.

The loading vector of the first principal component (explaining 79.7% of the total variance) is shown in Fig. 3a. This component is acting as a levelling component and is extracting information from the whole spectrum, and as such it does not differentiate very much between different frequencies. This behaviour is typical for a levelling/normalising component, where the extracted information is of a non-analytical nature such as different S/N ratios. The second principal component, explaining 9.5% of the importance of the describing variables, separates *Sphagnum fuscum* from *Carex rostrata* and is shown in Fig. 3b. The third loading vector, which describes the incubation time, is shown in Fig. 3c. Obviously there are some chemical differences between *Carex rostrata* and *Sphagnum fuscum*, which can be monitored with ^{13}C NMR, and further there are chemical changes with time. The loading

Table 3
Significance and explained variance of each principal component in each PC analysis performed. The total explained variance, the input data, the data pretreatment and the number of variables used in each analysis are also shown

Data set [a]	PC no. [b]	PRESS/SS [c]	Limit [d]	Explained variance (%) [e]
CALC1	PC1	0.2078	0.9828	79.7
NN, spc.	PC2	0.5393	0.9826	9.5
94 var.	PC3	0.9008	0.9825	1.4
148 obj.				**Σ90.6**
CALC2	PC1	0.6673	0.9830	33.9
N, spc.	PC2	0.8795	0.9828	8.9
94 var.	PC3	0.9151	0.9827	6.6
145 obj.				**Σ49.4**
CALC3	PC1	0.1652	0.9840	83.3
NN, spc.	PC2	0.3977	0.9838	9.8
205 var.	PC3	0.8428	0.9837	1.1
89 obj.				**Σ94.2**
CALC4	PC1	0.2167	0.9786	78.9
NN, spc	PC2	0.5549	0.9787	9.6
94 var	PC3	0.8881	0.9785	1.7
93 obj.				**Σ90.3**
CALC5	PC1	0.6768	0.9465	35.7
N, spc.	PC2	0.8792	0.9444	14.4
94 var.	PC3	0.8547	0.9421	11.7
23 obj.				**Σ61.8**
CALC6	PC1	0.3209	0.9677	68.9
NN, FID	PC2	0.3877	0.9672	19.4
62 var.	PC3	0.9450	0.9666	3.2
60 obj.				**Σ91.5**

[a] CALC(number) is referring to the different data sets in the text. N = normalised to constant area, NN = non-normalised, var. = number of variables used, type of input, data spc. = spectrum, FID = free induction decay, the number of variables and objects included.
[b] The principal component number. PC1 = the first principal component, etc.
[c] PRESS = prediction sum of squares, the squared differences between observed and predicted values for the data kept out of the model fitting in the CV procedure. SS = sum of squares, residual sum of squares of the previous dimension.
[d] LIMIT = the confidence (95%) limit, which PRESS/SS should not exceed in order to be considered a significant principal component.
[e] The explained variance of each principal component. The total explained variance of the PC's is shown in bold writing.

vector from the second principal component indicates chemical functionalities that differentiate between *Sphagnum fuscum* and *Carex rostrata*.

The aim of this paper is not to go in to details or speculations about chemical compounds formed and vanished during the incubation time. This will be dealt with in a following paper where calculations have been conducted in such a manner that these questions hopefully can be answered.

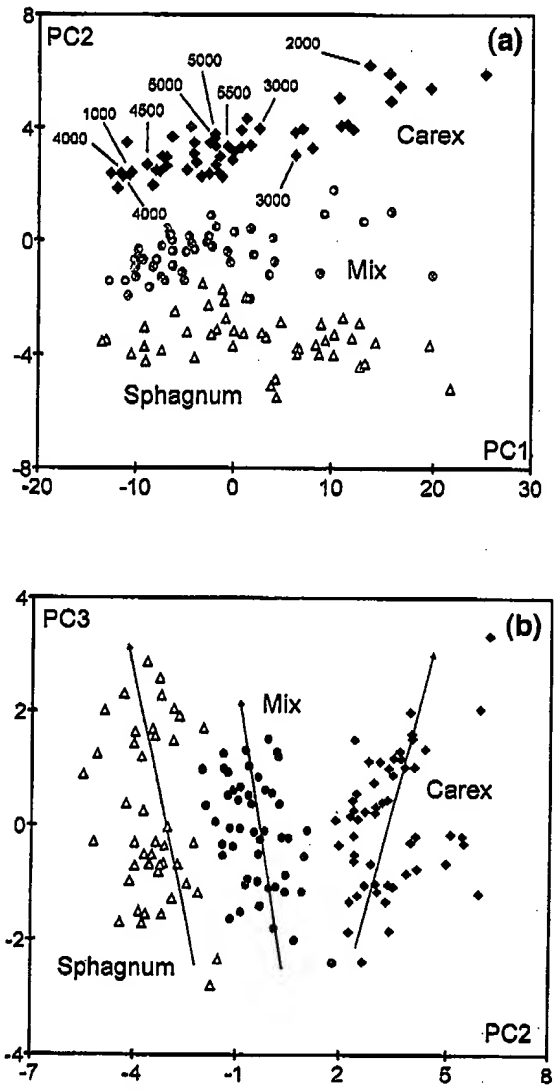


Fig. 2. (a) CALC1: 2nd versus 1st principal component using 94 non-normalised variables. Plant classes are marked *Sphagnum fuscum* = open triangles, *Carex rostrata* = black squares and mix (mixture 1:1) = grey circles. Carex objects marked with numbers (1000–5500) are identical samples except that they differ in the number of scans in the NMR experiment, e.g. S/N ratios are different. (b) CALC1: 3rd versus 2nd principal component, using 94 non-normalised variables. Plant classes are marked as in Fig. 2a and the arrows show the main decomposition direction for each plant class.

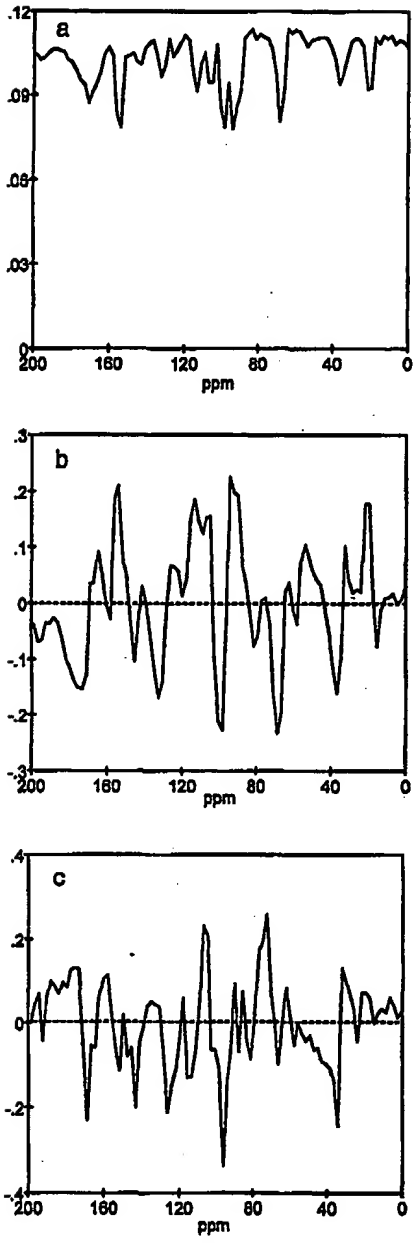


Fig. 3. (a) CALC1: Loading vector of the first principal component versus the chemical shift. 94 non-normalised variables were used. The loading vector uses the whole spectrum for information extraction. (b) CALC1: Loading vector of the second principal component versus the chemical shift. 94 non-normalised variables were used. (c) CALC1: Loading vector of the third principal component versus the chemical shift. 94 non-normalised variables were used.

3.2. Instrumental variations and drift

To examine the source of the pattern of objects explained by the first principal component one peat sample was subjected to a controlled variation of the different number of scans (NS) in the NMR experiment. This procedure produces a set of objects differing only in the S/N ratio. The objects in Fig. 2a marked with numbers ranging from 1000 to 5500 are from the same peat/plant sample (*Carex rostrata* of the 5th sampling occasion treated with non-flushing nitrogen, condition C in Table 1). The numbers indicate the NS used for each object. The objects with the same number of scans are clustered together, indicating a good stability of the NMR instrument. The relationship between the first principal component and objects with different S/N ratios is, however, far from simple, since no clear trends can be observed.

Thus there is an explanation to the non-systematic object pattern described by the first PC in Fig. 2a. The main variance in the data set (excluding the control objects with different number of scans) is of a non-analytical nature, where instrumental drift and variations are the probable sources.

There are probably several reasons for this non-analytical behaviour. Firstly, the Hartman–Hahn [38] condition in the CP/MAS experiment is set manually by turning a knob, which sets the power used for the carbon-13 excitation pulse, and visual determination whether the condition is set correctly. It is difficult to make this decision exactly repetitive every time. Secondly, due to different densities of the samples and the amount of available material, the actual analysed material can vary, thus causing variations in S/N ratios. Thirdly, the short and medium long term variations and drift in the spectrometer itself may be an additional source of variation. The short term drift/variation will be averaged out since each analysis time is approximately 3.5 hours long, but the medium long-term drift will affect each sample differently and thus introducing an instrumental variation within each sampling occasion.

There might also be an additional variation introduced by the spectrometer, a long-term variation due to fluctuations in the power output from the high power amplifier, instability of the proton decoupler unit, tuning of the CP/MAS probe and the high

power amplifier, variations in temperature, ageing of the electronic components and other electronical variations. This long-term variation could cause a systematic drift between each sampling occasion but it should appear as a non-analytical variation.

The fact that 79.7% of the variation in the data set is of a non-analytical nature and is of no interest could result in a total neglect of the first principal component. Alternatively, a normalisation step prior to the analysis could be applied.

3.3. The effect of normalising

There are several methods to use when normalising the data, for instance setting one of the peaks to a constant value. In this manner the variation of the chosen peak will be lost and the neighbour data points will be partially fixed. An internal standard compound could also be used to which normalising is referenced. Another and probably better method is to normalise each spectrum to a constant area. The objects that differ only in number of scans are spread over the whole PC, which indicates that the first PC is acting as a normalising component. If one would normalise data prior the analysis those objects would cluster together. In this manner the non-analytical variation, i.e. the S/N ratio, should be reduced considerably.

The non-normalised NMR spectra used in CALC1 were normalised to constant area and a new PC analysis was performed (CALC2). Three statistically significant principal components explained 49.4% of the total variance of this data set. In Fig. 4 the second (PC2) versus the first principal component (PC1) is shown. Further details on the calculation can be found in Table 3.

By using the normalising procedure the variation due to the different S/N ratios has been reduced considerably. This can be observed in Fig. 4 where the objects of the same sample with different number of scans are clustered together. A comparison between the result using normalised (Fig. 4) or non-normalised (Fig. 2a) data shows that the class separation is slightly better in the non-normalised case (Fig. 2a) but the objects within the classes are more disperse.

The first principal component in CALC2 (normalised data), explaining 33.9% of the total variance

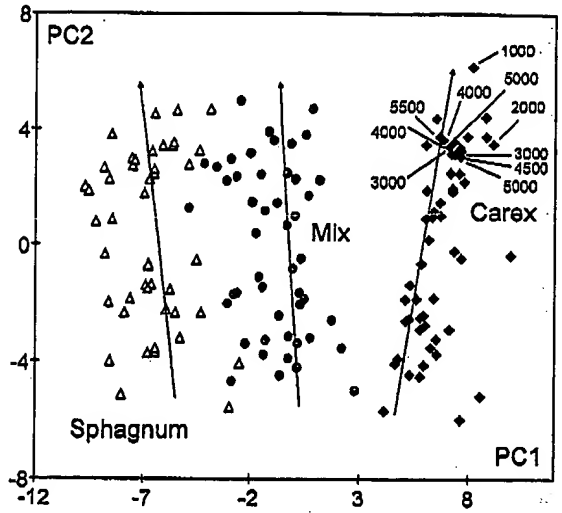


Fig. 4. CALC2: 2nd versus 1st principal component, using 94 normalised variables. Plant classes are marked as in Fig. 2a and the arrows show the main decomposition direction for each plant class. Carex objects marked with numbers (1000–5500) are identical samples except that they differ in the number of scans in the NMR experiment, e.g. S/N ratios are different.

resembles the second principal component in CALC1, explaining 9.5% of the total variance. Further, the second PC in CALC2, explaining 8.9%, shows a very similar object pattern as the third PC in CALC1, explaining 1.4% of the variance. When the data are normalised the total explained variance is low (49.4% in CALC2), which indicates that there might still be information to be explained. After the two first principal components the interpretation of the scores and the loadings becomes very complicated.

Since the object patterns in the score space are almost identical when using normalised and non-normalised data one would expect the subspectra to be very similar. This is also the case since the first loading vector of CALC2 is very similar to the second loading vector of CALC1. Further, the second loading vector of CALC2 also shows a similar pattern as the third loading vector of CALC1. However, there is one difference between the loading vectors, especially between the second loading vector of CALC1 and the first loading vector of CALC2. When normalised data are used the loading vector tend to display broader peaks, which can be seen in Fig. 5, which shows the first PC using normalised data compared to Fig. 3b where non-normalised data

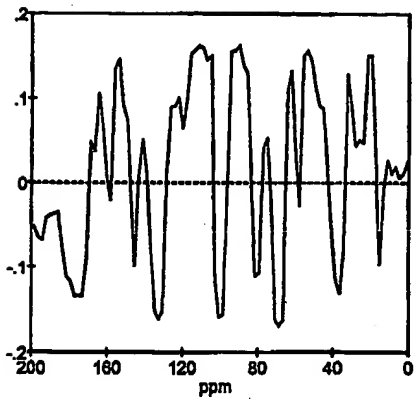


Fig. 5. CALC2: Loading vector of the first principal component versus the chemical shift. 94 normalised variables were used.

were used. This phenomenon is more obvious when 205 variables are used (not shown). In the normalising procedure the first levelling principal component of CALC1 (non-normalised data) has vanished. This supports the conclusions made earlier that the first PC in CALC1 describes variations that can be assigned to instrumental non-analytical variations such as different S/N ratios and this can be reduced considerably by the normalising procedure. The normalisation of the data causes a rotation of the data and the loadings and the scores are very similar to those in the non-normalised case excluding the first dimension.

Concluding remarks

In our case we will use non-normalized data, due to (1) the slightly improved class separation compared to the normalised data and (2) the sharper peaks in the loading vectors when using non-normalised data. Generally, normalising procedures should always be used with great care, since this pretreatment tends to obscure some of the systematic variation. It is quite possible that the first normalising principal component does not only describe the non-analytical variation, but also some of the interesting analytical variation obscured by the non-analytical variation.

3.4. Number of variables

When digitising a spectrum there is always the question of how many variables should be used in the multivariate data analysis. Nordén and Albano [18] pointed out that the number of variables (74 or 1025) was not very crucial in that study. However, it does depend on the relative size of the variations that are of interest. In this study we have used 94 or 205 variables to see whether there are any differences in the results depending on the number of variables used.

A PC analysis was performed using 205 variables (CALC3). Not all objects were included in this calculation, i.e. the objects from the 3rd and 5th sampling occasions were excluded, leaving 89 objects. The NMR spectra were not normalised by any means. Three significant principal components were extracted describing 94.2% of the total variance in the data set. Further details on the calculation can be found in Table 3.

To be able to compare the results from CALC3 with a calculation done on exactly the same data set but using 94 variables, a new PC analysis was performed (CALC4). Three significant principal components explained 90.2% of the total variance in the data set. Further details on the calculation can be found in Table 3. The results of CALC4 is very similar to the results in CALC1, the small differences could be assigned to the fact that a different number of objects was used in CALC1 and CALC4.

When comparing CALC3 and CALC4, the clustering of the objects in the score space is very similar. The different subspectra are also very similar between CALC3 and CALC4. According to the results from the scores and loadings output the use of 205 variables instead of 94 does not seem to increase the information.

However, the chosen number of variables can be crucial depending on the variation of interest. The next article in this series describes the problem of the relative chemical changes taking place during decomposition of plant material under different redox conditions. A separate PC analysis was performed on *Sphagnum fuscum* objects only, and to limit the variation even more only the air treated objects were included. Thus the number of objects was 12 and the number of variables was either 94 or 205. Fig. 6a and 6b show the two loadings of the principal components that describe the incubation time. Fig. 6a is the result of 94 variables in the calculation and Fig. 6b is based on 205 variables and in both cases the

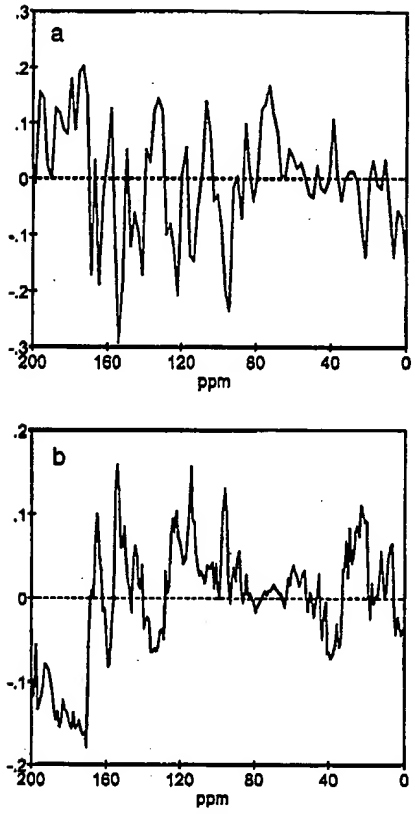


Fig. 6. (a) The peat class is *Sphagnum fuscum* which has been treated with air. The whole incubation series is included (12 objects) and the number of variables is 94. The figure shows the loading vector of the second principal component which describes the decomposition direction. (b) The peat class is *Sphagnum fuscum* which has been treated with air. The whole incubation series is included (12 objects) and the number of variables is 205. The figure shows the loading vector of the second principal component which describes the decomposition direction.

data used was non-normalised. The explained variances for the shown loading vectors were in the case of 94 variables 5.8% and for the 205 variables 11.1%.

In the case with 94 variables some of the peaks are described with only one data point, which is not satisfying. In the case with 205 variables the peaks are usually described with more than one data point. It should be mentioned that if one of these two loadings were inverted (positive peaks shifted to negative peaks) the direction of decomposition in the score space would be the same. In comparison with each other these two loading vectors have a similar pattern.

Conclusively, when choosing the number of variables we recommend the use of 205 (instead of 94), in order to be able to monitor small variations within the samples. Generally, one should choose as many variables as needed to avoid description of peaks with a single data point.

3.5. The effect of the LB parameter

When doing modern NMR experiments the received signal is a FID, which is the current induced by the evolving magnetisation which relaxes towards equilibrium, as detected by the receiver coil in the probe. A typical FID is shown in Fig. 7. The oscillation of the signal does in fact describe not one but a combination of several frequencies. The FID is mathematically transformed from its time domain to the frequency domain of a spectrum using Fourier transformation (FT).

The S/N ratio in the NMR spectrum can be enhanced by multiplying an exponential decaying function to the FID. This is possible because the actual analytical signal dominates in the first part of the FID (see Fig. 7), whereas the last part mostly consists of noise. It should be mentioned that the shown FID is just the first informative part (3 ms) of the collected FID, which does extend much further in time (35 ms). However, the fast decay of the signal can be seen also with this limited part of the FID

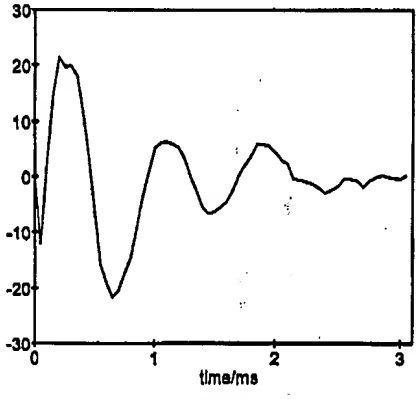


Fig. 7. FID (free inductive decay) of the original *Sphagnum fuscum* plant material. The figure is showing the signal intensity versus time.

(Fig. 7). By the exponential multiplication the first part of the FID is enhanced. By carefully adjusting the LB parameter in the exponential function it is possible to optimise the S/N ratio with a modest broadening of the signals. This can be done by choosing the LB value close to the line width at half height, which means that the spectrum has to be rather well resolved. If the LB value is chosen large there will be a smoothening of the spectrum. In our case the severe signal overlap causes a large variation in line width, nevertheless a line broadening is usually applied to the spectra even though the effect of it in the PC analysis is little known.

Fig. 1a and Fig. 8 show two NMR spectra of *Sphagnum fuscum* with LB = 0 Hz and LB = 200 Hz, respectively. The effect of the drastic increase in line width can clearly be seen. The S/N ratio increases dramatically and the fine structure is lost. How will this affect the result of the PC analysis?

To investigate the influence of a varying LB a PC analysis was performed (CALC5) on a data set of *Carex rostrata*. The objects were air-treated samples from the whole incubation period (i.e. 12 objects). One object from the third sampling occasion has been subjected to eleven different LB values ranging from zero to 100 Hz. The first two principal components, which explain 50.1% (35.7% + 14.4%) of the total variance in the data set, are plotted against each other in Fig. 9. A more detailed information about the statistics of the calculation is given in Table 3.

The LB variation is described by a combination of

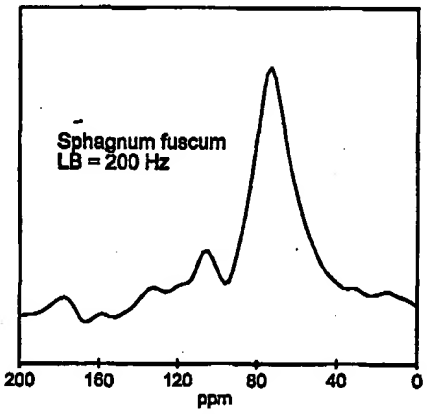


Fig. 8. ^{13}C CP/MAS NMR spectrum of the original *Sphagnum fuscum* plant material. Line broadening (LB) = 200 Hz.

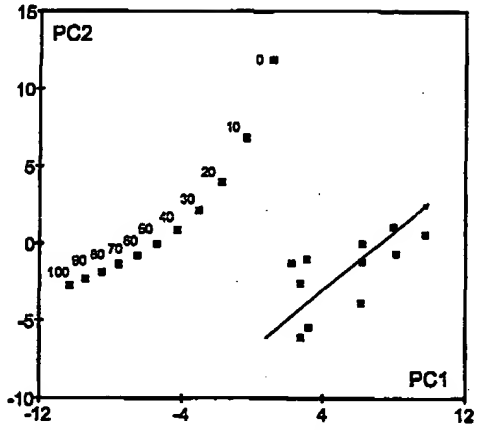


Fig. 9. CALC5: 2nd versus 1st principal component. The calculation is performed on air-treated *Carex rostrata* samples. One of the samples in the third sampling occasion has been treated with eleven different LB (line broadening) values (0–100 Hz). The smooth curve of objects are the chosen objects with different LB values. The arrow shows the main decomposition direction.

the first and the second principal component. Further it can be concluded that the variations caused by a differing LB is rather large compared to the variations caused by the chemical changes due to decomposition. However, our interpretation is that these two variations are causing different clusters which are detected by a combination of the first and second principal component. The average decomposition direction is marked with an arrow, which indicates increased humification. If the objects subjected to different LB were excluded in the calculation the PC's are tilted and describe the chemical change due to the humification and the plot looks quite different. This means that variations of the LB value cause new dimensions which are added to the model and which are quite different compared to the variation caused by the chemical information. A PC analysis which has been performed using one object from each botanical class, each subjected to 40 different LB values ranging from 0 to 1000 Hz, shows that the class separation information starts to diminish even at low LB values, i.e. 30–50 Hz.

Concluding remarks

The choice of LB in CP/MAS NMR is an important parameter since information starts to diminish even at low LB values. Since the effects are very